Automating the Anonymisation of Textual Corpora

Author: Laura García Sardiña

Advisors: Arantza del Pozo and Izaskun Aldezabal

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

September 2018

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Laburpena

Gaur egun, testu berriak etengabe sortzen doaz sare sozialetako mezu, osasun-txosten, dokumentu ofizial eta halakoen ondorioz. Hala ere, testuok informazio pertsonala baldin badute, ezin dira ikerkuntzarako edota beste helburutarako baliatu, baldin eta aldez aurretik ez badira anonimizatzen. Anonimizatze hori automatikoki egitea erronka handia da eta askotan hutsetik anotatutako domeinukako datuak behar dira, ez baita arrunta helburutzat ditugun testuinguruetarako anotatutako corpusak izatea. Hala, tesi honek bi helburu ditu: (i) Gaztelaniazko elkarrizketa espontaneoz osatutako corpus anonimizatu berri bat konpilatu eta eskura jartzea, eta (ii) sortutako baliabide hau ustiatzea informazio sentiberaren identifikazio-teknikak aztertzeko, helburu gisa dugun domeinuan testu etiketaturik izan gabe. Hala, lehenengo helburuari lotuta, ES-Port izeneko corpusa sortu dugu. Telekomunikazio-ekoizle batek ahoz laguntza teknikoa ematen duenean sortu diren 1170 elkarrizketa espontaneoek osatzen dute corpusa. Ordezkatze-tekniken bidez anonimizatu da, eta ondorioz emaitza testu irakurgarri eta naturala izan da. Hamaika anonimizazio-kategoria landu dira, eta baita hizkuntzakoak eta hizkuntzatik kanpokoak diren beste zenbait anonimizazio-fenomeno ere, hala nola, kode-aldaketa, barrea, errepikapena, ahoskatze okerrak, eta abar. Bigarren helburuari lotuta, berriz, anonimizatu beharreko informazio sentibera identifikatzeko, gordailuan oinarritutako Ikasketa Aktiboa erabili da, honek helburutzat baitu ahalik eta testu anotatu gutxienarekin sailkatzaile ahalik eta onena lortzea. Horretaz gain, emaitzak hobetzeko, eta abiapuntuko hautaketarako eta galderen hautaketarako estrategiak aztertzeko, Ezagutza Transferentzian oinarritutako teknikak ustiatu dira, aldez aurretik anotatuta zegoen corpus txiki bat oinarri hartuta. Emaitzek adierazi dute, lan honetan aukeratutako metodoak egokienak izan direla abiapuntuko hautaketa egiteko eta kontsulta-estrategia gisa iturri eta helburu sailkapenen zalantzak konbinatzeak Ikasketa Aktiboa hobetzen duela, ikaskuntza-kurba bizkorragoak eta sailkapen-errendimendu handiagoak lortuz iterazio gutxiagotan.

Abstract

A huge amount of new textual data are created day by day through social media posts, health records, official documents, and so on. However, if such resources contain personal data, they cannot be shared for research or other purposes without undergoing proper anonymisation. Automating such task is challenging and often requires labelling in-domain data from scratch, since anonymised annotated corpora for the target scenarios are rarely available. This thesis has two main objectives: (i) to compile and provide a new corpus in Spanish with annotated anonymised spontaneous dialogue data, and (ii) to exploit the newly provided resource to investigate techniques for automating the sensitive data identification task, in a setting where initially no annotated data from the target domain are available. Following such aims, first the ES-Port corpus is presented. It is a compilation of 1170 spontaneous spoken human-human dialogues from calls to the technical support service of a telecommunications provider. The corpus has been anonymised using the substitution technique, which implies the result is a readable natural text, and it contains annotations of eleven different anonymisation categories, as well as some linguistic and extra-linguistic phenomena annotations like code-switching, laughter, repetitions, mispronunciations, and so on. Next, the compiled corpus is used to investigate automatic sensitive data identification within a pool-based Active Learning framework, whose aim is to obtain the best possible classifier having to annotate as little data as possible. In order to improve such setting, Knowledge Transfer techniques from another small available anonymisation annotated corpus are explored for seed selection and query selection strategies. Results show that using the proposed seed selection methods obtain the best seeds on which to initialise the base learner's training, and that combining source and target classifiers' uncertainties as query strategy improves the Active Learning process, deriving in steeper learning curves and reaching top classifier performance in fewer iterations.

Contents

List of Figures

List of Tables

1 Introduction

A growing amount of new data —and especially textual data— is created every day through social network posts, health records, pieces of news, official documents, and more. These data often contain personal information and, although they could be valuable and beneficial for research, transparency, or other purposes, they may be kept unshared due to the prohibitively high costs of their manual anonymisation and to the potential legal repercussions if it is not done correctly. For this reason, the automation of textual anonymisation is a task that has gained increasing attention in recent years.

In Natural Language Processing (NLP), lack of annotated data is a common issue when using Machine Learning (ML) approaches, especially if the desired language for the task is a low-resourced one or, in some cases, a language other than English. In the data de-identification scenario, little anonymised and annotated corpora are available to train robust classifiers in supervised settings. Even if there are a few labelled anonymised corpora available, these are often not directly exploitable for that purpose, since they are anonymised through techniques which do not preserve readability and result in unnatural texts. Such is the case of corpora like the Dortmund Chat Corpus in German (Lüngen et al., 2017) or the French SMS corpus (Panckhurst, 2013). In this context, using Active Learning (AL) can help optimising the annotation process, resulting in better classifiers trained with fewer annotated data. Another solution to annotated data scarcity is to use Knowledge Transfer (KT) techniques, which make use of existing annotated resources to improve learning in new tasks, domains and/or languages. Thus, combining KT with AL to re-use information from different available annotated corpora for more efficient textual anonymisation can be considered a sensible solution.

The present thesis has two main objectives: (i) to compile a new annotated anonymised corpus in Spanish, and (ii) to exploit that corpus to investigate the automation of the anonymisation process in cases where little annotated data are available.

For the first objective, work has been done mostly manually. The full corpus compilation procedure has been completed from the transcription of raw audio data, including some annotations of linguistic and extra-linguistic phenomena, to the annotated anonymisation of the data. The anonymisation process itself has involved several steps: selection of the anonymisation technique, definition of the cases that need to undergo anonymisation, identification of those items in the corpus, categorisation of the identified elements, replacement of the elements considering the selected technique, and revision of the resulting anonymised data.

For the second objective, the focus has been set on automating the identification phase of the anonymisation process, which has been approached as a binary classification task where the positive label corresponds to the elements that refer to sensitive information and so need to be de-identified. The proposed approach consists of training a classifier in an already available anonymisation annotated corpus, even if it differs in language and domain from our target corpus, and to use that classifier's uncertainty over the target unlabelled corpus as KT technique to improve two key aspects of AL in the target domain: seed selection and query selection strategy. The aim has been to train a robust classifier for the

target data taking advantage of an existing labelled corpus, and having to label as little instances as possible, thus speeding up the process. Results on a simulated AL scenario show that the proposed approach does obtain better seeds on which to initialise the learner and that the selected queries help reach top classifier performance in fewer iterations.

The main contributions of this thesis are the following:

- ES-Port, a new anonymised spontaneous spoken human-human dialogue corpus in Spanish has been compiled, including annotations regarding the anonymised items and their categories, as well as some linguistic and extra-linguistic phenomena like code-switching, laughter, repetitions, or misspellings. Such corpus has been made publicly available for research purposes.

- A new approach is proposed to improve seed selection and query selection strategy in Active Learning by exploiting Knowledge Transfer. The presented methods use uncertainty from a transferred source classifier trained on an annotated source corpus to select the best possible seed from an unlabelled target corpus. The AL process is further improved by combining uncertainty information from the source and target classifiers as query selection strategy, which speeds up the base learner's learning curve in the target domain. The approach achieves good performance even when the source and target corpora differ both in language and content domain.

- A strong baseline is set for the anonymisation task using the ES-Port corpus. Also, to the extent of my knowledge, the anonymisation task is tested for the first time within the Active Learning framework.

This thesis is structured as follows: first background information about data anonymisation and Active Learning is provided in Section 2; then the methodology to be followed in this work is displayed in Section 3; next, the full compilation process of the ES-Port corpus is reported in Section 4, including its anonymisation and some notes on the resulting resource; Section 5 explores new methods for seed and query selection in a simulated Active Learning setting where the previously compiled corpus is used as target in a sensitive data identification task. Finally, in Section 6 some conclusions and directions for future work are given. Most of the work included in this thesis appears in (García-Sardiña et al., 2018a,b), attached in Appendix II (page 51).

2 Background

This section introduces the data anonymisation task (2.1) , which is the task at hand in this thesis, and the Active Learning framework (2.2), which is the one followed to approach annotated data scarcity. For the former, the focus is set on textual anonymisation, providing information about existing practices and standards and different approaches to the task. For the latter, the framework is explained in detail and different approaches to its key aspects are presented, as well as some previous works addressing its combination with Knowledge Transfer.

2.1 Data Anonymisation

Data anonymisation consists of treating personal data in such a way that, once the procedure is finished, the data can no longer be used to identify any natural person (*data subject*), neither directly nor indirectly, while the value and usefulness of the original data is preserved. Data protection laws, like the European General Data Protection Regulation (GDPR) (Council of European Union, 2016) among others around the globe, demand datasets to be adequately anonymised before sharing —or selling— them to be exploited for other purposes.

There are mainly three key risks present in anonymisation which any effective technique should prevent, listed and described below:

1. Singling out: the possibility to isolate the data relating to some distinguishable individual among the rest.

2. Data linkability: the possibility to link identifiers referring to the same individual or group within the same source or from other sources.

3. Inference: the possibility to infer a link between two or more values in the data with significant probability.

2.1.1 Anonymisation Techniques and Standards

Several techniques have been designed that try to successfully anonymise data and avoid identification risks. As described in Article 29 of the European Data Protection Working Party (Council of European Union, 2014), predecessor of the new GDPR, anonymisation techniques typically fall within two main broad types:

- *Randomisation*: family of techniques that involve data alteration so the link between the data and the data subject is removed without losing its value. The techniques included in this approach are: (i) noise addition, adding random small changes in the data; (ii) permutation, swapping certain attribute values across individuals, and (iii) differential privacy, inserting more or less noise considering a specific query by a particular third party while the original data are preserved.

- *Generalisation*: family of techniques which involve diluting the data subjects' attributes or reducing their granularity. There are two subgroups of generalisation techniques: (i) aggregation and k-anonymity, generalising attribute values in such a way that data subjects are grouped with at least other k individuals; and (ii) the l-diversity technique, which extends k-anonymity ensuring that every attribute in an equivalence class has at least l values, and the t-closeness technique, which refines l-diversity requiring the values to mirror the original distribution of the data.

However, these methods are considerably oriented to the anonymisation of structured datasets in the form of graphs or tables, like databases with attribute-value slots (e.g. "age": "54", "gender": "female", "diagnosis": "flu"). A better suited classification of techniques specifically oriented to the anonymisation of unstructured textual data is exposed in works by Medlock (2006) and Dias (2016):

- *Suppression* or *removal*: a neutral place-holder replaces the item to be anonymised, e.g. "XXXX", "ANON".
- *Tagging* or *categorisation*: a label indicating its category or identifier is used to replace the item to be anonymised, e.g. "LOC", "LOCATION453".
- *Substitution* or *pseudonymisation*: the item to be anonymised is substituted by one of the same category (e.g. "Michael" can be a substitute for "Joseph"). There are several ways in which this substitute choice can be done: the new item can be randomly extracted from a dictionary, 'intelligently' substituted by an item that shares the same features (e.g. item with property "male person first name" by one of same characteristics), swapped with another item to be anonymised too within the text document, or obtained by applying some degree of generalisation to the item (e.g. "Great Wall Restaurant" generalised to "Restaurant").

2.1.2 Steps in the Textual Anonymisation Process

Considering a typical textual anonymisation scenario, the process usually involves the following steps:

1. **Selection** of the anonymisation technique (Section 2.1.1) to be employed taking into account the desired result.
2. **Definition** of what types of elements need to be anonymised. The types of elements may vary considering the data's domain: the items to be anonymised in a medical record may not be exactly the same as the ones in a company's contract with a client. However, there can probably be some cases which could count as domain-independent and be considered target of anonymisation across datasets (e.g. person names).
3. **Identification** of items falling into the given definition of elements to be anonymised in the target data.

4. **Categorisation** of the elements. This step can be combined in the identification phase or even completely omitted if the selected anonymisation technique does not require it (as is the case of suppression).

5. **Replacement** of the identified items by some neutralising elements which will depend on the selected anonymisation technique (e.g. their category label, some random pseudonym, an invariant label like 'anon').

6. **Revision** of the obtained result, since treating sensitive data should always require a double check of the output.

Some of the mentioned steps in anonymisation can be fully or partially automated, while others should not. The identification and categorisation steps are easily automatable if annotated data on which to train a classifier are available. If data with the needed annotations are not available, it would be desired that a model could be optimally trained without having to annotate most of or even the full dataset. This is the aim of Active Learning. Also, if there are some annotated data available for the target task, but the type of annotations or domain of the data do not exactly match the needs of our target, it would be interesting to see if those data can still be exploited somehow to help in the classifier's training. Knowledge Transfer can help with this resolution.

2.1.3 Approaches to Textual Anonymisation in NLP

Data anonymisation is a complex task which involves a high need of adaptation to every target dataset and difficult decision-making to what constitutes a sensitive element that needs to be anonymised and what does not. Typically textual anonymisation has been carried out manually. Medlock (2006) anonymised a corpus of emails following two different guidelines with different degrees of exhaustiveness stating what constitutes a sensitive item and what does not, which shows the subjectivity entailed in the task. Manually anonymising a dataset, even if not too large, is a tedious, time-consuming, and highly expensive task. Dorr et al. (2006) reports the numbers of manually de-identifying medical notes containing Personal Health Information (PHI) items: each note containing an average of 7.9 target sensitive items took around 87.3 seconds to complete.

Due principally to the prohibitively high time and monetary expenses of manual anonymisation there has been an increasing number of attempts in the research community to automate, either fully or partially, textual anonymisation in recent years.

Many of such attempts have taken place in the medical domain, where the target is to de-identify Electronic Health Records (EHR), usually for their use in research, but attempts at automating textual anonymisation have taken place in diverse domains, like the legal or the social media domains, and using a wide variety of approaches.

Tveit et al. (2004) propose an approach where unigrams (tokens) are matched against different in-domain compiled dictionaries corresponding to different categories and substituted by pseudonyms. In a similar fashion, Patel et al. (2013) simply check every word in the text against dictionaries of items that need to be anonymised and items which do not,

for an SMS corpus anonymisation task. In the legal domain, Bick and Barreiro (2015) use pattern matches and contextual rules to identify personally identifiable information in a parallel English-Portuguese legal corpus. Using a more statistical approach for anonymising posts in the social media domain, Nguyen-Son et al. (2015) check the co-occurrence of noun phrase chunk pairs in the posted text context against a compiled large dataset and decide to anonymise the chunks if the co-occurrence score is above a certain threshold, e.g. the score of the pair (Shinjuku, Tokyo) will be high in comparison with that of the pair (Shinjuku, Harvard University), so the former will be considered sensitive while the latter will not.

A common approach to the anonymisation task is to address it as an Named Entity Recognition and Classification (NERC) problem. So is the case of Kokkinakis and Thurin (2007), who constructed a rule-based system to identify sensitive items and then removed them. Dias and colleagues (Dias, 2016; Dias et al., 2016) implement an anonymisation system for text documents which first uses a named entity recognition module and then applies co-reference resolution for the entities found to cover all references to personal data in the text.

Other approaches to the automation of textual anonymisation have also applied Machine Learning (ML) techniques to build classifiers to be integrated in the anonymisation process. Medlock (2006) trains a HMM-based classifier on his manually anonymised email corpus to recognise sensitive data and set a baseline for others. In the medical domain, Szarvas et al. (2007) extracted different types of features from the records (local context, regular expressions, dictionaries) and applied ML to construct a binary classifier to identify each word as PHI or non-PHI. Li and Qin (2017) use local features (e.g. length, part-of-speech), global features (term's position in the document), and external features (e.g. term appears in some gazetteer, proper nouns list, medical concept lexicon) to categorise terms according of their identification risk and tag them using ML ensemble classifiers. Depending on its classification, each term can be removed or substituted (high risk, explicit identifiers like name, surname), anonymised using a cluster-based method (medium risk, quasi-identifiers like date of birth, hospital, age), or be left unchanged (irrelevant and low risk, health and medical details like symptoms, medications).

Although as has been mentioned above there have been some approaches to face anonymisation as a classification task using Machine Learning, up until this point and to the extent of my knowledge Active Learning had never been applied to the textual anonymisation task to speed up training such classifiers.

2.2 Active Learning

Active Learning (AL) is an approach in ML to overcome the problem of annotated data scarcity. In AL the target is to train the best possible classifier using the minimum annotated data, thus reducing time and annotation costs.

There are mainly three possible scenarios for AL considering the specific ways in which the classifier being trained may ask queries, i.e. ask for annotations of some instance(s): *Query Synthesis*, *Stream-Based Selective Sampling*, and *Pool-Based Sampling*. While the

Algorithm 1 Pool-Based Active Learning typical setting

Input: set of labelled instances L, pool of unlabelled instances U, query strategy ϕ, batch size B, stopping criterion S
repeat
// Train model M on L
Q = best set in U of size=B according to ϕ
// Ask Oracle to label Q
$L = L + Q$
$U = U$ - Q
until S or size(U)=0
return M, L

first two have been generally overlooked in the literature, Pool-Based Sampling has gained most attention and is the most common and explored scenario in AL. It is also the selected one in this work. For these reasons, the rest of this thesis will refer to that AL scenario only.

The basic terminology utilised in the Active Learning literature includes the following concepts:

- *Seed*: a small set of initially labelled instances.
- *Pool*: the set of target unannotated instances, initially large.
- *Base Learner*: model.
- *Oracle*: labelling source, typically human annotator.
- *Query Selection Strategy*: criterion to select the best instance(s).
- *Stopping Criterion*: criterion to make the Active Learning stop.

In typical pool-based AL scenarios, the input elements are the seed and the pool. The annotated seed is used to train the base learner, which then asks the oracle to label the instance (in the case of *serial* AL) or set of instances (in the case of *batch mode* AL) which it considers more informative according to the chosen query selection strategy. The newly labelled data are moved from the pool to the labelled set and the model is retrained following this process iteratively until the stopping criterion is satisfied or the pool is empty. The output obtained is a classifier and a set of newly annotated data. The overall process is shown in Algorithm 1.

There are several questions to be addressed in Active Learning research, mainly: how to select the seed on which to start annotating and learning, which query selection strategy will be best to speed up the classifier's learning curve, what is a good criterion to make the classifier stop learning, and which is a good metric for cost evaluation. These issues and how they have been approached in the literature are further explored in the following subsections.

2.2.1 Seed Selection

The issue of seed selection has not received much attention in the literature, but there have been several approaches to get the best initial labelled set to speed learning from iteration zero.

Olsson (2008) explored the importance of a good seed selection in AL by comparing the use of a random seed against using cluster-centroid based document selection in a NERC annotation task. However, results showed little to no improvement over random sampling at seed point.

In an extensive survey of AL in NLP, Olsson (2009) claims that the seed should be representative of the classes to be handled, defending that:

> "[...] Omitting a class from the initial seed set might result in trouble further down the road when the learner fits the classes it knows of with the unlabeled data it sees. Instances that would have been informative to the learner can go unnoticed simply because the learner, when selecting informative instances, treat instances from several classes as if they belong to one and the same class."

However, the author does not provide any empirical data supporting his claim. Also, it must be noticed that in a real case scenario the true class labels of the instances would not be known.

Tomanek et al. (2007) compare multiple kinds of seeds in a NERC task for the medical domain: a tuned seed, selected after checking all sentences of the data against entity gazetteers manually created by specialists and ranking them from largest to smallest number of diverse mentions in them and selecting the top ones; a second seed containing only sentences with no mentions to any of the entities in the compiled gazetteer; and a random seed. They report significant improvement in using the tuned seed over the other two. Continuing with this work, Tomanek et al. (2009) compare four types of seeds, again in a NERC task, this time considering majority and minority classes: a random seed; a majority seed, containing sentences with at least one majority class entity, and no minority class entities; a minority seed, containing a large number of minority entity classes; and an outside seed, containing no entity mentions. They reported best results with the minority seed, while using majority and outside seed performed worse than random. This gives support to the claim made by Olsson (2009) mentioned before that class representativeness in the seed set is important.

Dligach and Palmer (2009, 2011) propose to use unsupervised language model (LM) sampling by training a LM on the unlabelled corpus and selecting as seed those instances with the lowest probability in a word sense disambiguation (WSD) task. The expectation of this method is to select more instances of the rare classes than using random selection. They obtain better results using this method than using a random seed.

In this thesis we test and compare different seed selection methods to determine which of them are actually beneficial for initialising the Active Learning.

2.2.2 Types of Query Selection Strategies

The core of any AL approach lays on the selected query strategy. Several query selection criteria have been designed and put into practice in the AL literature with the aim of finding the most informative instances in the pool data. As surveyed by Settles (2010, 2012), there are six main types of serial query selection strategies:

- Uncertainty sampling (Lewis and Gale, 1994): the learner uses some uncertainty measure (like a label's predicted posterior probability or entropy) to query the instances about which it is less certain how to label. However, this approach may be prone to query outliers.

- Query-by-Committee (QBC) (Seung et al., 1992): a committee of models present different hypotheses and query the instances on which they disagree the most. Disagreement can be measured in different ways: vote comparison, vote entropy, Kullback-Liebler divergence... Nonetheless, this type of strategy may also be prone to query outliers.

- Expected model change: the instances queried are those expected to cause the greatest change to the current model if their label was know. A strategy in this framework is using Expected Gradient Length (EGL) with discriminative models (Settles et al., 2008b), but it is computationally expensive and also tends to query outliers.

- Expected error reduction: the model estimates the expected future error of the instances in the unlabelled pool and queries those with the minimal expected risk (Roy and McCallum, 2001). Approaches in this framework include minimising the expected 0/1-loss or log-loss, maximising the expected information gain of the query or the expected mutual information of the output. As was the case of using EGL, this type of strategy is also computationally expensive.

- Variance reduction: the learner queries the instances which minimise the output variance and thus the model's generalisation error, i.e., those instances which minimise the learner's inverse Fisher information or the Fisher Information ratio (Zhang and Oles, 2000).

- Density-weighted methods (Baker and McCallum, 1998; Settles and Craven, 2008; Settles, 2008): in this setting the best queries are those which are both uncertain to the model and representative of the data's underlying distribution.

However, in practice querying and retraining a learner in serial may be too time-consuming, expensive, and overall inadequate for real case scenarios. Batch-mode seems like a more efficient option in such cases. In batch-mode AL the challenge is to select the optimal set of instances to query. Simply selecting the top queries according to some serial query strategy probably will not result in the optimal batch, since information overlap inside the set is not taken into account. In order to find the best batch Shen et al. (2004)

propose three key notions to be considered: *informativeness* (related to uncertainty), *representativeness* (of the data distribution), and *diversity* (avoiding information overlap). Some approaches to finding the best batch include the following:

- Brinker (2003) incorporates diversity among batches with Support Vector Machines (SVMs) by selecting instances with maximal angles with respect to each other. This approach is computationally cheap and scalable.
- Hoi et al. (2006a,b) use Fisher Information to create batches that are both diverse and informative.
- Xu et al. (2007), also using SVMs, incorporate density by querying centroids of clusters lying closest to the decision boundary.
- Guo and Schuurmans (2008) vectorise the pool of unlabelled instances and try to find the most informative batch directly by using gradient search.
- Chen et al. (2017) combine LDA clustering and uncertainty sampling to informative and diverse batches.

Although the query selection strategy to be used should be decided considering if the AL will be applied in serial or in batch mode, in practice it is common to test and report results for serial query strategies using batches (choosing the top instances). This is reasonable considering time and cost efficiency issues, but it should be acknowledged that such approach is not the optimal one.

In this work, batch mode AL is used in combination with uncertainty sampling, which is one of the most commonly —and successfully— used query selection methods in the literature.

2.2.3 Stopping Criterion

In AL scenarios a point is often reached when the cost of continuing to query an oracle is higher than the cost of the errors the model may make. This often happens when the model has accomplished top performance or its learning curve has reached a plateau. A good stopping criterion should make the iterative process stop at an optimum or near-optimum point.

Several stopping criteria have been proposed in AL research. Some of them have the drawbacks of being dependent on a specific type of model or query selection strategy (e.g. hyperplane-based criteria are only usable in scenarios where the chosen model set-up uses a hyperplane-based geometrical algorithm, query-by-committee agreement-based criteria can only be used in those AL scenarios), others of requiring an external annotated validation set, which may not be optimal considering that AL tries to overcome annotated data scarcity making use of as little labelled data as possible. A few, however, do not have such drawbacks and are applicable in any AL setting. In this section some of the proposed stopping criteria are presented.

Schohn and Cohn (2000) use AL with SVMs, labelling examples that lie closest to the dividing hyperplane. They propose a heuristic stopping rule that labelling should stop when all the instances in the SVM margin have been labelled. When the best query candidate is not closer to the hyperplane than the support vectors, the authors understand the margin has been exhausted and the process should be terminated. Another hyperplane-based stopping criterion is introduced by Fu and Yang (2015), who propose to stop AL when the SVM separating hyperplane lies in a sparse region of the feature space, based on the idea that its generalisation error has reached a local minimum then.

Zhu and colleagues (Zhu and Hovy, 2007; Zhu et al., 2008a,b, 2010) use AL for WSD and text classification tasks and they propose several stopping criteria: *Max-Confidence* (also called *Maximum Uncertainty*) , *Min-Error* (also called *Selected Accuracy*), *Minimum Expected Error Strategy*, *Overall-Uncertainty*, and *Classification-Change* (or *Threshold Update Strategy*). Max-Confidence is based on uncertainty: the process should stop when the entropy score of the selected unlabelled instances is close to zero. Min-Error relies on the oracle's feedback on whether the current classifier is predicting the labels correctly, similar to a human annotator. Minimum Expected Error Strategy uses estimations of the classifier's error on future instances and terminates the process when this error is as low as possible. Overall-Uncertainty is similar to Max-Confidence, except it takes into consideration the entropy of all the instances in the unlabelled pool. Classification-Change stops AL when the predicted label for each instance in the pool does not change for two consecutive iterations.

Vlachos (2008) uses the classifier's confidence in an uncertainty sampling strategy setting for text classification, NERC and shallow parsing tasks. The proposed criterion terminates AL when the model's confidence on an external validation set becomes constant or drops for some number of consecutive iterations. Also using a validation set, Bloodgood and Vijay-Shanker (2009) propose stabilising predictions as stopping criterion for AL. The idea is to consider the predictions made on a separate unannotated validation set and stop the AL when they have stabilised, even if it is not known whether such predictions are correct or not. Another criterion based on variance is that proposed by Ghayoomi (2010), who defends that once a classifier is trained, the variability of its confidence scores on the unlabelled data and, therefore, its variance start to decrease. Based on this assumption Ghayoomi proposes to stop the process at the point when variance has reached its peak and starts to decrease.

Laws and Schütze (2008) report their results on using three new stopping criteria in uncertainty sampling AL for a NERC task: *Minimal Absolute Performance*, *Maximum Possible Performance*, and *Convergence*. The first involves setting a minimal threshold before the AL begins. The process is terminated when the classifier's performance reaches this minimal threshold. The second stops the classifier from learning when it reaches its optimal performance given the data. The last uses the gradient of the learner's estimated performance or uncertainty to terminate AL when the available data stop contributing to the improvement of the model's performance.

Olsson and Tomanek (2009) propose the *Intrinsic Stopping Criterion* (ISC), which combines the notions of Selection Agreement (Tomanek et al., 2007), SA, and Validation

Set Agreement (Tomanek and Hahn, 2008), VSA. This criterion is oriented to Query By Committee settings and it proposes to stop the AL process when the SA, i.e. the decision of the next selected query from the pool, is greater than or equal to the VSA, i.e. decision on an unannotated validation set.

Of the several questions which can be addressed in AL research, this thesis focuses on two: seed and query selection. Since finding a good stopping criterion is not one of the key aims in this work, the decision for stopping the AL process in the experimentation phase is simply when the classifier reaches top performance. The top performance score is known since the AL experiments are simulated. Contributions to this question for a real use case scenario would require further research on this topic.

2.2.4 Cost Evaluation

In most AL research it is assumed that reducing the number of instances to be labelled directly translates into reducing the cost. However, this reduction in the number of labelled samples does not necessarily imply cost savings. For this reason, several proposals on how to better measure cost have been developed in the literature. Here, a few are presented.

Culotta and colleagues (Culotta and McCallum, 2005; Culotta et al., 2006) take into account not only the number of instances that need to be labelled, but also their annotation difficulty. Difficulty is measured as the number of edits needed given partially labelled instances, using the predictions made by the current model.

Kapoor et al. (2007) use a cost metric which includes both the cost of acquiring a new label, as the expected information value for learning that label, and the costs of the expected model's misclassification errors, assuming the cost of a label is linear in the instance's length. Also related to length, Tomanek et al. (2007) use the number of tokens of the queried sequences in the labelled set to measure annotation costs.

In a parsing task, Hwa (2004) uses the number of constituents that need to be labelled in each instance to estimate cost. However, in a similar task, Baldridge and Osborne (2004) claim that the cost of labelling each instance is more related to the number of possible parses than to the number of constituents, what they call *discriminant cost* (versus *unit cost*, which is the number of instances in the labelled set). On the other hand, Haertel et al. (2008) uses the estimated time needed to annotate each sequence as cost measure, also in sentence parsing. Using time as cost measure was also studied by Settles et al. (2008a), who determined that even though annotation time is variable across instances and annotators, it can still be accurately predicted after seeing only a few examples.

As it has already been mentioned, AL research in this thesis focuses on the seed and query selection questions. Although cost evaluation is an important aspect in AL, it is not researched deeply in the current work. Instead, the assumption is made that more queries imply more annotation time and so more expense. Therefore, the number of instances queried is used as cost metric. Replicating the presented experiments taking into account different cost measures could be interesting as future work.

2.2.5 Knowledge Transfer for Active Learning

Knowledge Transfer (KT) or Transfer Learning comprises the idea of re-using existing annotated data, referred to as *source*, to improve the learning of other data, called *target*, which may differ from the former in domain or task. Knowledge Transfer is a very wide field and it encompasses different types of settings, such as multi-task learning, domain adaptation, sample detection bias, or co-variate shift. A complete survey for KT is gathered in (Pan et al., 2010). There have been several works in the NLP field which combine Knowledge Transfer with Active Learning to face annotated data scarcity. In this section, a few are shortly reported.

In a sentiment analysis task, Rai et al. (2010) use an unsupervised domain adaptation technique to learn an initialiser distribution separator hyperplane. Then they use distances to that hyperplane to select and query the most divergent instances from both the source and target data, updating the separating hypothesis in each iteration. However, relying on a hyperplane narrows down the possible classification algorithms to be applied, plus it may not be suitable for sequential data (Álvarez et al., 2017).

Chattopadhyay et al. (2013) perform domain adaptation from a source to a target domain using source data weights, and then querying the instances in the target domain which are most similar according to their marginal probabilities.

A somewhat similar approach is that taken by Huang and Chen (2016). In a scenario where both the source and the target domains have too little labelled data and an oracle is only available for the former, they apply AL in the source domain and then perform domain adaptation from the labelled source data to the target data. This approach differs from the previous two in that it applies domain adaptation *after* the AL process and not before nor during it, therefore not having any effect on it.

Some studies have used classifiers trained on source data for AL before. Shi et al. (2008) train a classifier on the source domain data plus a small set of annotated target data instances and use it to answer the queries made by the base learner in the target domain. In their setting, the oracle is only asked to label instances when the prediction confidences are too low.

Saha et al. (2011) perform unsupervised domain adaptation on the target data, adapting the source data representation to make its marginal distributions more similar to those of the target data, using a *domain separator* which exploits the relatedness of the two domains. In a similar fashion to Shi et al. (2008)'s work, their approach includes a hybrid oracle which uses a *free oracle*, i.e. a classifier trained on the source data, to label the queried target instances which are related enough to the source data, and a traditional manual annotator which they call *expensive oracle* to label the queried target instances which are not related enough to the source data.

In a multi-domain sentiment classification task, Li et al. (2013) train a source classifier on the labelled source data and a target classifier on a small annotated set of the target data. Then they use a QBC strategy to select the most informative samples based on prediction disagreement and make a labelling classification decision based on a graph-based label propagation algorithm.

Recently, Shao (2018) integrated KT into the classical query-by-committee AL strategy by including different models trained on the source and target domains, and assigning different weights to each committee member in every iteration.

As has been seen, several studies have proposed methods to combine KT and AL to face annotated data scarcity. Many of such studies have integrated domain adaptation and exploited the domains' similarity or divergence for or after the Active Learning process. Others are more similar to the approach in this thesis and make use of classifiers trained on source data, either to work as a labelling oracle or to be part of a committee of classifiers for query selection. As opposed to the latter, which use disagreement between source and target classifiers to select queries, the proposal in this thesis is to use the classifiers' uncertainty on the unlabelled data, independently of whether they agree or not, for ranking and selecting the best queries.

3 Methodology

This section presents the methodology followed in this work to reach two objectives: first, the compilation of a new corpus from scratch, anonymised and including specific anonymisation category labels (Section 4), and second, the exploitation of such corpus to explore a combination of Knowledge Transfer with Active Learning in a simulated new —initially unannotated— data anonymisation task (Section 5).

For the corpus creation phase, the compilation involved transcribing raw audio data, including some linguistic and extra-linguistic annotations during transcription, analysing the type of contents of the data, and anonymising them for their secure sharing. The selected anonymisation technique was that of Substitution or Pseudonymisation, so as to maintain the readability, cohesion, and naturalness of the data while preserving privacy. The definition step was unavoidably done manually, carefully considering the type of data contained in the corpus itself, as well as taking into account the linguistic intricacies of the given language and potential inconsistencies after token substitution. Since there were no annotated data from which to depart, the identification step was also done manually, with the help of an interface designed for that purpose. Although this was also the case for categorisation, some attempts at automation were made to help speed up this phase. Replacement was a fully automated process given gazetteers for the different categories, which were gathered manually. Revision is necessarily a human-made, manual process. The full corpus compilation process is presented visually in Figure 1. In Section 4 the different steps and the issues faced during the full process are presented and discussed in detail.

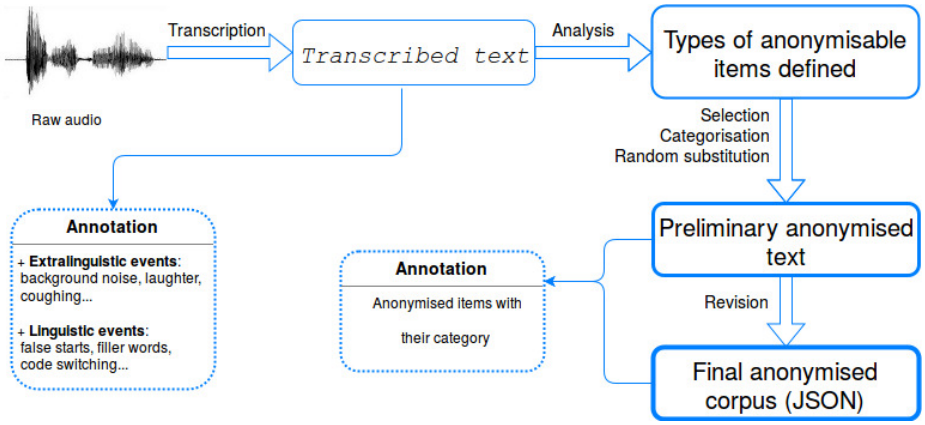

Figure 1: ES-Port corpus compilation process.

In the second phase of this work the newly created corpus is used to explore improved automation of one key step in the anonymisation process: identification of the items in the data that need to be anonymised. This step can be seen as a binary classification task, where the positive label should be assigned to those items which refer to personal data and so need to undergo de-identification. The question in this phase is how to build a classifier when no annotated data from the target domain are available, while having

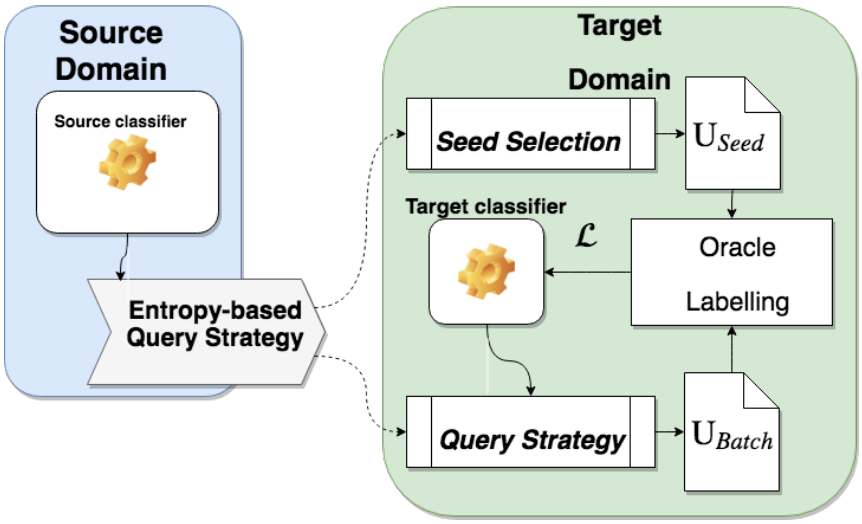

Figure 2: Knowledge Transfer scenario: knowledge from a classifier trained on source domain data is used first for seed selection and then for query strategy in the target domain, where the Active Learning is applied.

at hand some (even if small) amount of annotated data for the same task but from a different language and/or domain. The proposed approach consists of training a source classifier on the source annotated corpus and use knowledge from that transferred model, specifically its uncertainty on how to label the target unannotated corpus instances, to select an optimal seed set on which to start Active Learning. Once AL is initialised a target model is iteratively trained on batches of instances of the target data, and the source model's and the target model's uncertainty scores are combined and used as query selection strategy. This approach is compared with several other methodologies in evaluation, like K-Means nearest centroid or largest utterance selection. Query by uncertainty has obtained satisfactory results in AL using a wide range of algorithms for the base learner: logistic regression (Lewis and Gale, 1994), Support Vector Machines (Schohn and Cohn, 2000; Shen et al., 2004), Hidden Markov Models (Scheffer et al., 2001), Conditional Random Fields (Culotta and McCallum, 2005; Settles and Craven, 2008), and more. Here, considering the sequential nature of the task at hand, discriminative models based on Condition Random Fields (Lafferty et al., 2001) —CRFs— are used. The Knowledge Transfer for Active Learning scenario proposed is represented visually in Figure 2. Such approach is presented and discussed in further detail in Section 5.

In order to evaluate the performance of the seed selection strategies, the F1 score on the target test set obtained by the base learner trained using the selected seeds is reported. Being *Precision* the number of true positive results divided by the number of true positive plus false positive results returned by the classifier, and being *Recall* the number of true positive results divided by the number of true positive plus false negative results, the F1

score is calculated as follows:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

For measuring the performance of the elected query selection strategies in Active Learning, the most common, straight-forward practice is to observe the base learner's learning curve as it gets iteratively trained. Such curve is plotted setting some classification metric (like accuracy, area under the curve, or, like in this work, the F1 score) in the Y axis and the number of instances labelled and used for training in the X axis. In each training iteration a point is plotted where the number of instances used and the achieved score meet, resulting in a curve which depicts the classifier's learning. If the AL is working successfully, this curve will be much steeper than using passive learning —the more, the better—, and ideally it will reach top performance before the pool is exhausted. Assuming less labelled instances implies less needed annotation time and expenses, best AL results will be those with the lesser number of labelled instances needed to reach a satisfactory performance level.

4 Compiling the ES-Port Corpus

In this chapter the full compilation process of the ES-Port corpus is presented. ES-Port is a collection of 1170 spontaneous spoken human-human dialogues from calls to the 24/7 technical customer support service of a telecommunications operator for companies, which provided the raw audio data. The resulting corpus includes linguistic and extralinguistic phenomena annotations as well as anonymised sensitive items with their corresponding labels. The work described in this section has been published in (García-Sardiña et al., 2018a), as part of the 11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018, Miyazaki (Japan). The original paper is attached in Appendix II.

4.1 Compilation Process

As displayed in Figure 1, the full corpus compilation process has consisted of four main steps: transcription, analysis, anonymisation and revision. Each of these steps is explained in detail in the following subsections.

4.1.1 Audio Transcription

The first step in the compilation process has been the transcription of the raw audio data into text. The raw recorded calls contain both speech and non-speech sounds, such as channel-associated noises, laughter or background music. Due to its spontaneous character, the type of speech used is more focused on conveying the message than on taking care of form itself. As a result, it includes phenomena like mispronunciations, false starts, repetitions, non standard forms, segment overlap between speakers (interruptions), etc. The complete audio dataset consists of 40 hours, with an average length of around two minutes per conversation.

Automating the transcription task using generic large vocabulary continuous speech recognition (LVCSR) systems has proved to be not feasible. Given the challenging spontaneous nature of the telephone recordings, word error rates (WER) of their automatic transcriptions were too high (i.e. 77.21% on a test set) to save transcription times.

Eventually the transcription process has been carried out fully manually using the Transcriber 1.5 annotation tool (Barras et al., 2001). The tool guidelines have been followed for transcription. Considering this and regarding punctuation, even if the norm in Spanish is to use opening question (¿) and exclamation (¡) marks, only closing ones are transcribed in this corpus. Quotation marks are also not included. Instead, when a speaker is citing or reading something, the first word in the citation is capitalised as if it were the beginning of a new sentence. Other transcribed punctuation marks are comma (,), dot (.) and apostrophe (’), although the latter is not used in Spanish and rarely occurs in the corpus. Also, numbers have been transcribed in their full word form, e.g. *quince* (fifteen), *doscientos cuarenta y tres* (two hundred forty three).

Apart form the orthographic transcriptions themselves, the following linguistic and extralinguistic phenomena have also been annotated:

- Language events: indicate a switch to a language different to Spanish by the speaker. These switches are only annotated when the speaker pronounces the word(s) correctly in the target language, otherwise they are left unannotated (if pronounced as if they were Spanish) or annotated with a mispronunciation event (if pronounced freely). Languages other than Spanish in ES-Port are: Basque, Catalan, Asturian, French, Italian and English, the latter being the most common one (up to 91.59% of language event occurrences).

- Pronunciation, noise and silence events: annotated following the conventions of the Transcriber tool. Some examples of these are: whispered pronunciation, segment unintelligible, mouth noise, electric noise, breathing, cough or throat cleaning, whistling, background laughter, music, etc.

- Other annotations: unfinished words and nonwords, repetitions and false starts, lengthening in pronunciation, typical Spanish shortenings in pronunciation – such as using *pa* instead of *para* or dropping of intervocalic *d* in final syllables (e.g. *demasiao** for *demasiado*, *entrao** for *entrado*)–, and some continuers and filler words (e.g. *o sea*, *eh*, *hala*, *mhm*, *ay*).

The result of this first phase is the transcribed text, including the mentioned linguistic and extralinguistic annotations.

4.1.2 Anonymisation Oriented Analysis

The second step in the corpus compilation process has involved analysing the data, in order to identify which types of elements in it may refer to sensitive information or could be used in any possible way to identify the individuals involved, endangering their right to confidentiality. Given that the conversations in the corpus correspond to an IT Customer Support setting, the dialogues in it often contain multiple sensitive data sharing between the clients and the technicians, including domain specific cases pertaining digital trace data. Considering this, the types of elements that have been decided to anonymise are the following:

- Individuals' basic personal information elements: names, surnames, name diminutives or nicknames, personal identification numbers.

- Contact information and digital trace elements: phone numbers[1], user names/numbers, postal addresses, email addresses, IP addresses, web domains.

- Workplace and organisation-related elements: names of organisations, NIFs (tax identification number) and CIFs (tax code), easily linkable names of products and services, prices.

[1]Some prefixes were kept if relevant to the conversation.

	NER			NERC		
	Pr	Rc	F1	Pr	Rc	F1
IXA Pipes	0.32	0.62	0.42	0.26	0.50	0.34
FreeLing	0.36	0.65	0.47	0.24	0.54	0.33
CoreNLP	0.47	0.96	0.63	0.36	0.99	0.53

Table 1: NER and NERC Precision (Pr), Recall (Rc), and F1 scores using three different available taggers on test set.

- Other elements: locations, dates, card numbers and bank accounts, trouble ticket numbers, dispatch notes, passwords, spellings of any of the previous elements.

The result of the second phase is the scheme above, defining the types of elements in the corpus data selected to undergo anonymisation.

4.1.3 Anonymisation: Selection, Categorisation and Substitution

The third step in the compilation process has involved finding the items falling into the scheme defined in the previous step and anonymising them. The chosen anonymisation method was anonymisation by substitution, since the result is a natural text. The anonymisation process has consisted of three phases: (i) selecting the items to be anonymised in the corpus, (ii) categorising them, and (iii) substituting them for items of the same category. The full anonymisation process has been carried out at token level in a semi-automatic way.

Initially, attempts were made to automate the first two phases trying several Named Entity Recognition and Classification (NERC) tools available for Spanish: IXA Pipes (Agerri et al., 2014), FreeLing (Carreras et al., 2004), and Stanford CoreNLP (Manning et al., 2014). Even though these taggers have achieved good results on planned written language such as news texts datasets, the spontaneous nature of the ES-Port corpus' data poses an extra challenge. As it can be seen in Table 1, running some trials on a small test set of the target corpus gave results too poor to automate the selection and categorisation phases of the items to be anonymised, since their application would still require much manual revision and correction, producing no time savings. It is important to note that even if the results of the NERC tools had been better, their output would still not have been enough to cover all the types of items identified in our anonymisation scheme.

In the end, the selection phase has been completed manually with the help of an interface developed for that purpose. With this tool, each dialogue text is loaded on screen on selection by dialogue name, preserving turn structure. Each token to be anonymised is selected with a double click, and when the dialogue is completely anonymised it is saved in a structured JSON format.

Next, those tokens identified as sensitive need to be categorised for their proper substitution. Considering the types of elements selected for anonymisation, the following categories have been defined in order to classify them:

- Female name
- Male name
- Surname
- Organisation
- Place
- Month
- Mail
- Domain
- Letter
- Product/Service
- Number I (0-9)
- Number II (10-29)
- Number III (tens 30-90)
- Number IV (hundreds)

Since anonymisation is approached on a token level basis, these categories can account for all the types of items agreed to undergo anonymisation in Subsection 4.1.2, for example Spanish personal IDs consist of a sequence of numbers plus a letter, prices and phone numbers are straight-forwardly accounted for by the Number categories, first names and nicknames can take any of the first two categories depending of the person's gender, while user names and passwords may be a combination of letters, numbers, names, and more.

Accordingly, gazetteers for each of the categories have been created[2] for the substitution phase, which has consisted of random substitution with an item of the same category. The reason why there are for different categories for numbers is to keep coherence in the resulting anonymised text after token substitution. The four categories have been defined taking into consideration Spanish full word number writing norms and irregularities. Let us see some cases illustrating why numbers were not to share one unique gazetteer: *Cuarenta* (40, type II) can be combined with *cuatro* (4, type I) to form *cuarenta y cuatro* (44), but *diez* (ten, type III) cannot (*diez y cuatro** is not a valid option in Spanish). At the same time, *cuatrocientos* (400, type IV) can be combined with type II (e.g. *cuatrocientos veinticuatro*, 424), type III (e.g. *cuatrocientos cuarenta*, 440), type I (e.g. *cuatrocientos cuatro*, 404), typed III and I combinations (e.g. *cuatrocientos cuarenta y cuatro*, 444), or be left alone. With this division, substitutions by numbers of the same type category ensure coherence keeping the same number of tokens as the original data.

With sensitive items selected, the categories defined, and the gazetteers compiled, the categorisation and substitution phases have been unified using a program developed for that purpose. The program shows each token to be anonymised according to the selection phase and asks for its category label, which is introduced manually. The token is then substituted by a randomly chosen item from the gazetteer corresponding to its category. Once an item is anonymised, its substituting element is kept throughout the whole dialogue so as to keep coherence, but not across dialogues in order to avoid the risk of linkability.

The result of the third phase is a preliminarily anonymised corpus in structured JSON format, now including information about anonymised tokens and their category. Although the different Number categories come in handy for gazetteer-based substitution, in the labels kept in the resulting JSON these are unified to a unique Number category, which has a more universal character and overcomes language specific conventions.

[2]Items in the organisation and domain gazetteers are made up and none of them existed at the time the anonymisation was carried out.

4.1.4 Revision

The final step has consisted in manually revising the resulting corpus. This is an important step which should never be ignored when dealing with sensitive data, independently of whether anonymisation is carried out manually or automatically.

The main purpose of revision is granting that no personal data has been overlooked in anonymisation and so it can safely be shared without risking the right to privacy of the individuals involved. In the ES-Port corpus compilation scenario, this step has also included the aim of ensuring that the resulting text is consistent and coherent. The most common corrections which have had to be done in revision include the following:

- **Names' short forms:** proper names and their short forms were not considered the same reference during substitution, so their resulting anonymisation was not coherent. As an example, imagine in a conversion one of the participants is called *Francisco*, but he is often referred to as *Fran*. If *Francisco* is replaced by, let us say, *Pedro* and *Fran* is replaced by *David*, for example, the result is not natural, since *David* is not short for *Pedro* nor the other way around and the person involved would be switching names, which is more than unusual.

- **Names' cross-linguistic equivalences:** although this issue was not too common, there were several cases where participants gave a name in some language and their equivalent in Spanish to facilitate comprehension, for example "*Jaume, como Jaime pero en catalán*". As in the previous case, random conversion of these names resulted in unnatural situations and had to be manually corrected and changed to some equivalent cases.

- **Compound names:** names consisting of at least two parts could result in unnatural combinations after random substitution. For example, *José Manuel*, *Juan María*, *María José*, or *María Luisa* are valid person names in Spanish, but *Héctor Iñaki**, *Jaime Helena**, *Marta Juan**, or *Carla Paloma** are combinations improbable to happen in an everyday scenario. The same happened with some place names, being quite common those formed by "San" or "Santa" plus a person name: for example, *San Sebastián* and *Santa Marina* are real Spanish place names, but *San Omar** and *Santa Noelia** are not.

- **Date inconsistencies:** when discussing dates some inconsistencies were found in cases where a date was mentioned in reference to another. These inconsistencies occurred after substitution in cases originally like "*estamos en mayo, el mes pasado fue abril*" ("it's May, last month was April"), "*hoy es lunes 25, el viernes es 29*" ("today it's Monday 25th, Friday will be the 29th"), "*lo mandé el diez del nueve, diez de septiembre*" ("I sent it on 9/10, September the tenth").

- **Number combinations with zero in unit position:** in the previous step (Subsection 4.1.3) number zero was included in the gazetteers for category Number type I (0-9). This produces some inconsistencies when combined with numbers of type III

Num. Dialogues	1170	Vocabulary size (types)	11.2k
Num. Turns	65.2k	Vocabulary Freq. (tokens)	535k
Avg. Turns per Dialogue	55.76	Filler Words (types)	37
Avg. Turns Length	8.20	Filler Words Freq. (tokens)	26.5k
Num. Overlap Turns	11.3k	Foreign Words Freq. (tokens)	3.3k
Noise/Pronun./Silence Events	11.5k	Language Switch Events	3k

Table 2: Statistics (rounded) of the ES-Port corpus.

and IV, e.g. *treinta y cero**, ("thirty zero"*), *doscientos cero** ("two thousand and zero"*). However not including zero in the units gazetteers would have resulted in a corpus with no appearances of such number (not even as uncombined units), which is not a real scenario and would have biased the data. Therefore these combination cases were corrected manually during revision.

- **Spellings:** during anonymisation each individual spelled letter was substituted at random by another without taking into consideration what is being spelled. For example consider a case in which a name and its spelling have being randomly substituted like this: "*Laura, L, A, U, R, A*" → "*Victoria, T, C, E, K, C*". As can be observed, the spelling does not match the target spelled word, and even the number of letters spelled may not be the same as the length of the target word. Random substitution of individual letters may be very helpful in cases like password spelling anonymisation, but in cases like spelling of names, surnames, domains, and so on it needs to be reviewed and modified to keep coherence.

The result of the fourth and last step is a coherent anonymised and annotated text corpus in JSON format, ready to be used and shared. Even if the corrections regarding coherence made in this last step are not strictly necessary to get an anonymised text, they are important to obtain a natural, realistic, and reusable result. If one achieves this result, an outsider person will not be able to distinguish whether a token has been anonymised or not without looking at its label, granting extra security if some token has been left non-anonymised by mistake.

4.2 The Compiled Corpus

Below, the resulting compiled corpus is described both quantitatively and qualitatively. A sample excerpt from a real dialogue is presented and its annotations are described. Also, some potential applications of the compiled corpus are stated.

4.2.1 Corpus Statistics

In Table 2 some statistics of the corpus regarding number of dialogues, turns, overlaps, vocabulary, fillers, and events are presented.

The numbers show typical attributes representative of spontaneous spoken conversation present in the corpus. For example, it includes a high number of overlapping turns (approximately 17% of the total turns) and a rich use of continuers and filler words (around 5% of the total vocabulary tokens). It is imperative to mention that since the dialogues are set in an IT domain scenario words in English appear quite often, making up to 91.59% of all non-Spanish words tokens and reaching up to 3.24% of the vocabulary.

4.2.2 Sample Data Description

To illustrate the type of annotations and data gathered in the ES-Port corpus, an excerpt from one of the actual dialogues included in the corpus is presented in Table 3. It does not show all the information included in the JSON-formatted corpus, but the table displays the necessary data to see real cases of events, overlaps, and other linguistic annotations together with the transcribed dialogue.

As has been said before, different types of speech and non-speech events appear often throughout the corpus. The excerpt shows such instances in turns 35 and 37, where the event tag of type "ch" indicates that the utterance was pronounced in a whisper, in turn 45, whose event tag says the segment was unintelligible and could not be correctly transcribed, and in turns 31 and 33, where the tag indicates that the speaker laughed before continuing talking.

Overlapping speech is a common phenomenon in the corpus. It occurs up to three times in this small excerpt: in turns 29, 41, and 43. Such overlap is indicated by the fact that two different speakers intervene within the same turn. It must be noted that an overlapping turn can contain speech, extralinguistic phenomena, or a combination of both.

False starts, repetitions, incomplete words, and nonwords are common in spoken spontaneous conversation. In the corpus, false starts and repetitions are tagged simply using `<word>`, like the repetition of *se* in turn 30. On the other hand, incomplete words and other nonwords are tagged as `<nonword->`, as shown in turns 31 and 45.

As was stated in Subsection 4.1.1, small deviations from what is supposed to be the standard pronunciation of some words also take place regularly along the corpus. An example of shortened pronunciation occurs in turn 30, where the word *enviado* is marked with a + symbol, indicating that the intervocalic phoneme /d/ in a final syllable is dropped in pronunciation. On the other hand, examples of lengthening in pronunciation occur in turns 30 and 40, characterised with a = symbol.

The excerpt also displays several instances of filler words (e.g. *o sea*, 'I mean') and continuers (e.g. *mhm*) which are frequently used in the corpus. These are marked following the pattern `<%word>`, with spaces removed when the item consists of more than one word. In the excerpt shown some of these cases can be seen in turns 31, 36, 38, 43, 45, and 46. Finding so many of them in such a small fragment gives an idea of their high frequency in spontaneous dialogue.

Finally, a language switch event takes place in turn 45. The annotation includes a short code for the target language of the segment, in this case English ('en'), which is the most usual one in the ES-Port corpus.

T	S	U
29	spk1	De todas formas
	spk2	esto
30	spk1	si has +enviado el correo estate tranquilo porque <se=> se para.
31	spk2	(*EVENT*: noise-rire)
	spk2	<%mm> Es <lom-> <i-> incluso si lo <envi-> <%aver> <su-> supuestamente hasta las cinco y media, no?
32	spk1	Sí.
33	spk2	(*EVENT*: noise-rire)
	spk2	Y si lo envío a las cinco y diez se cancela?
34	spk1	Sí, sí.
35	spk2	Ay, dios (*EVENT*: pronounce-ch)
36	spk1	<%mhm>
37	spk2	Ay, mi madre (*EVENT*: pronounce-ch)
	spk2	no puedo largarme de aquí digamos.
38	spk1	<%eh> si quieres
	spk2	<%eh>
39	spk1	llamar un poquillo más tarde y te intento pasar con él de nuevo.
40	spk2	Es que no hay ninguna forma de que, ningún número que yo pueda=
41	spk1	No, no.
	spk2	llamarles
42	spk2	a ellos o
43	spk1	No, <%osea>
	spk2	algo?
44	spk1	que le estoy llamando yo y no me responde.
45	spk2	(*EVENT*: noise-nontrans)
	spk2	<ueh->
	spk2	okay (*LANGUAGE*: en)
	spk2	<%pues> muchas gracias.
46	spk1	<%venga> a ti.

Table 3: Sample excerpt from a dialogue showing turn index (T), speakers (S), and the utterance segments transcription (U) including all the annotated information.

4.2.3 Potential Applications of the Corpus

The ES-Port corpus is a resource of real spontaneous spoken human-human dialogues in the IT technical support domain in Spanish, publicly available for research via META-SHARE[3]. In this subsection a few of the potential applications for which this resource could be valuable are mentioned.

One clear application of this corpus is its use to develop more open and natural dialogue systems in Spanish, and specifically in customer support settings. Even though it does not yet include dialogue act annotations, the corpus as it is could be used for research in

[3]The corpus can be found in the repository of the University of the Basque Country UPV/EHU `http://aholab.ehu.es/metashare/repository/search/`, under the name *ES-PORT*.

unsupervised approaches to dialogue system development. Such approaches would include language modelling to generate more human-like system responses and to detect and take into consideration the nuances of human-human spontaneous interaction, and to analyse and exploit turn-taking dynamics for incremental dialogue processing.

The corpus can also be exploited to develop NERC tools that work better for dialogue and spontaneous language in Spanish. The available anonymisation annotations can be generalised to account for the usual four (Organisation, Person, Location, and Miscellanea), six (plus Date and Number) classes or even some more, to train a NERC system for spontaneous language.

ES-Port can also be used in linguistic research in Spanish to study a wide range of issues. Some of these include: the use of discourse markers and filler words in conversation, their meaning in context, their effects and influence on the dialogue, code switching in the presented domain, and the strategies used for turn-taking and self-correction in spontaneous phone dialogue, among others.

Finally, the corpus can be exploited to explore the automation of text anonymisation and develop systems for that purpose. Automatic textual data anonymisation is the task at hand for which ES-Port is used in Section 5.

5 Textual Anonymisation using Active Learning with Knowledge Transfer

The motivation of the work corresponding to this chapter is to explore new ways to speed up the process of training a robust classifier for new textual data anonymisation by exploiting existing resources within the Active Learning framework.

In short, the idea is to train a source classifier on an available annotated anonymisation corpus and to use its uncertainty over the pool of unlabelled data from the target corpus first to select the best seed on which the initialise the Active Learning, and second to use that information together with the newly trained based learner's uncertainty as query selection strategy.

The work presented in this section will be published in (García-Sardiña et al., 2018b), as part of the 6$^{\text{th}}$ International Conference on Statistical Language and Speech Processing, October 15-16, Mons (Belgium). A preview of the original paper is attached in Appendix II.

5.1 Used Seed and Query Selection Methods

In this chapter, the different seed and query selection methods used in this work's experimentation are presented and described formally. First different scorers to measure sequence entropy are defined, then the methods based on K-Means Centroids and its combination with entropy are described, and finally the proposed approach to use entropy from a transferred source classifier for seed and query selection is introduced.

5.1.1 Entropy Scorers

Being $I = (w_1, w_2, ...w_{|I|})$ an instance (i.e., a sentence or utterance) composed of words of a corpus, the uncertainty over the binary decision for each word $w_i \in I$ can be measured using the Shannon entropy (Shannon, 1948):

$$H(w_i) = -P(\hat{y}_i = A|\ I)log_2(P(\hat{y}_i = A|\ I)) - (1 - P(\hat{y}_i = A|\ I))log_2(1 - P(\hat{y}_i = A|\ I))$$

where $P(\hat{y}_i = A \mid I)$ is the probability of the classifier assigning the *anon* label A to the word w_i. As each instance is a sequence of words, there are multiple ways in which the entropy score of the whole instance can be defined. In this work the following four types of sequence entropy scorers are used:

1. **H Sum:** The entropy value of the full instance is the sum of all its word entropies: $H(I) = \sum_{w \in I} H(w)$

2. **H Mean:** The entropy value of the full sequence is the mean of its word entropies: $H(I) = \frac{1}{|I|} \sum_{w \in I} H(w)$

3. **H K-Max:** The entropy value of the full instance corresponds to the mean of its K-Max word entropies: $H(I) = \frac{1}{K}\sum_{i=0}^{K} H(w)$, where the K words with highest entropy of the instance I are chosen.

4. **H Max:** The entropy value of the full instance is the maximum entropy value in the sequence: $H(I) = max_{w \in I} H(w)$

Entropy scorers can be used to measure how certain a classifier is about a certain labelling decision, resulting in a robust strategy to select the instances with highest information content —those about which the classifier is less certain, thus containing information still unknown to it— in the Active Learning process.

5.1.2 K-Means-Centroids Scorer

The K-Means clustering algorithm (MacQueen et al., 1967) can be used to split the pool data into K clusters or groups. Then, for each cluster, the closest candidate to its centroid is selected. Being B the batch size, i.e. the number of instances to select from the pool, let $K = B$ in the clustering algorithm, splitting the pool in B clusters. Then, being $c_1, c_2, \cdots, c_B$ the centroids of each cluster and I_{c_k} the instances that encompass the cluster of centroid c_k, the closest instance I_k to the cluster centroid according to the Euclidean distance is chosen for each cluster:

$$I_k = argmin_{I \in I_{c_k}} ||c_k - I|| \; \forall \; k = 1, \cdots, B$$

5.1.3 K-Means-Centroids-Entropy Scorers

Since the K-Means algorithm measures the input diversity and the $H(I)$ entropy scorers measure the base learner's uncertainty about an instance, both measures can be combined to select the instance I_k for each cluster:

$$I_k = argmin_{I \in I_{c_k}} ||c_k - I|| \cdot (1 - rescale(H(I))) \quad (1)$$

where the $H(I)$ results are rescaled so they fall within the range [0, 1].

5.1.4 Entropy-based Knowledge Transfer

As explained in Section 3 and as depicted in Figure 2, in this work it is proposed that the entropy measures from the source classifier (S-H Sum/Mean/K-Max/Max) can be exploited for both seed and query selection strategies in the target domain.

For seed selection, one cannot rely on knowledge from the target classifier (base learner) since in a realistic setting there are no initially labelled data in the target domain on which to train it. To overcome this limitation, source classifier's entropy score S-H (or its combination with K-Means using Equation 1) can be used to sample the most uncertain instances ($\mathcal{U}_{seed}$) in the target data as seed. After annotation of the selected seed, those labelled instances $\mathcal{L}(\mathcal{U}_{seed})$ can be used to start training the base learner and so initialise

the AL process. In addition, source entropy scorers (S-H) can be combined with the target classifier's ones (T-H) as query selection strategy and so select the next batch $\mathcal{U}_{batch}$ of instances to ask the oracle to label. The source and target combinations tested in this work are obtained by applying the dot product of the values obtained by different source and target entropy scorers.

5.2 Corpora

To test the proposed seed and query selection methods in a simulated AL setting, at least two anonymised and annotated corpora are needed: one source and one target. Apart from ES-Port, the corpus compiled previously in Section 4, ITAC (Medlock, 2006) is the corpus that has been used for that purpose, since it complies with the required criteria and is publicly available.

ITAC, the Informal Text Anonymisation Corpus, consists of about 2500 personal emails, both private and corporate, written in English and collected over seven years. Due to the nature of the data, errors and spelling, punctuation and capitalisation inconsistencies appear often in the dataset.

The ITAC corpus is anonymised by pseudonymisation and includes binary annotations (anonymised or not). It is already pre-partitioned into training, development and test sets, consisting of 666138, 6026 and 31926 tokens, respectively. Unfortunately, only the last two are annotated. Following the solution given in Medlock (2006) to the unannotated training set issue, the development set is used as training set instead and referred to as such.

Given the subjectivity of what constitutes a sensitive item that needs to be anonymised, there are two versions of ITAC annotated following two different schemes: a comprehensive one where each reference which might possibly be related to people or organisations is anonymised even if the risk of identification is very low, called *blanket* anonymisation, and a more selective one where only those references which directly refer to people or organisations are annotated, referred to as *selective* anonymisation. In the latter, items anonymised include person and organisation names and descriptors, postal addresses, phone/fax numbers, product names, user names, passwords and transactional identification and reference codes. Apart from these, the former also anonymises commercial titles, job titles, geographic and ethnic terms, and titles of films, books, TV shows, academic papers, courses and conferences. For the tests presented here only the blanket data is used, since in preliminary tests the best classifier was obtained using that set (results for the positive 'anon' label on their respective test sets were F1=0.80 using the blanket dataset and F1=0.73 using the selective dataset).

As opposed to ITAC, the ES-Port corpus is not pre-partitioned, so training, development and test sets can be chosen freely. In the case presented here, the first 900 dialogues (around 47k utterances after removing turns not containing any text, such as those where there is only laughter or unintelligible speech) have been taken as training set and the rest (23% of the data approximately) has been left as test set.

The main characteristics of the two corpora can be consulted in Table 4.

Characteristics	**ITAC**	**ES-Port**
Main Language	English	Spanish
Language Switching	No	Yes
Language Form & Style	Written, planned (emails)	Spoken, spontaneous (phone calls)
Content Domain	Various (personal, corporate)	IT, Telecommunications
Training Utterances	473	47073∗

Table 4: Comparison of the main characteristics of the ITAC and ES-Port corpora. ∗Number of utterances in ES-Port's training set after removing turns not containing any text; before that, total number of utterances was 50161.

5.3 Feature Selection

Since the proposed approach is tested on cross-lingual data, two source classifiers have been trained on ITAC for knowledge transfer: one with language independent features only and one including language dependent features as well. The target classifier has been trained on ES-Port using the language dependent feature set only. Features for each instance (i.e. sentence or utterance) result of the concatenation of the word-level features for each token in the sequence.

The word-level features used to train the language independent source classifier are the following:

- Position: [boolean] is Beginning Of Sentence (BOS), is End Of Sentence (EOS).
- Punctuation: [boolean] is comma, is dot, is question mark, is exclamation mark.
- Case / form: [boolean] is upper, is title, is digit.
- Tags: [string] Named Entity Recognition and Classification (NERC) tags and Part of Speech (PoS) tags.

For the language dependent source and target models, features are those included in the language independent feature set plus the following:

- Lexical: [string] word itself (lower cased).
- Morphological: [string] word's prefixes (first two and three characters) and suffixes (last two and three characters).

In both cases features also include all of the selected features in a [-2, +2] words context window. NERC and PoS tags were automatically extracted using the Stanford CoreNLP tool (Manning et al., 2014), which is available both for Spanish and English. However, the tag sets used by such tool for the two languages are different, and so they have been normalised to share the same values cross-lingually. In the case of NERC, tags correspond to the same entities types even if different names are used for them (e.g. Spanish tag 'LUG'

and English tag 'LOCATION' both refer to place entities), so their normalisation has only implied mapping each tag name to a new name in a one-to-one correspondence setting. In the case of PoS tags, on the other hand, the correspondence between the Spanish and the English tags is not straightforward. A new tag set consisting of sixteen PoS tags has been constructed and the outputted results by Stanford CoreNLP have been mapped to the new set in order to create uniform tag feature values across languages. More information about the normalised PoS tag set can be found in Appendix I (page 49).

5.4 Experiments

As advanced in the previous section, the ITAC corpus has been used as source data and the ES-Port corpus as target data for the experimentation phase. Thus, in the presented experiments, knowledge transfer has only been done from ITAC to ES-Port and not in the opposite direction. This is so because the annotated "training" set in ITAC is too small (473 utterances only) for its implementation in a realistic AL setting. Considering this, ITAC is referred to as *Source* and ES-Port as *Target* throughout the whole experiments section. All results are reported on the ES-Port test set and correspond to the positive ('anon') label only.

A CRF classifier has been trained passively on the whole ES-Port training data to test its top performance, reaching 0.935 of F1 score on the positive label. Source classifiers' top performance has also been tested on the ITAC test set, achieving a 0.803 F1 score using language dependent features and a 0.785 F1 score using language independent features.

Next, the experiment settings and results for seed selection are given in Section 5.4.1 and then several query selection strategies based on the proposed approach are tested and compared in Section 5.4.2, using the best seed obtained in the previous step.

5.4.1 Seed Selection Evaluation

The following methods have been implemented for seed selection:

- **Random:** the seed is selected at random. This is used as a weak baseline.
- **Maximum Utterance Length:** the samples with largest number of words are chosen as seed.
- **K-Means-Centroids (K-MC):** the K-Means algorithm is used to split the corpus and choose a representative sample in each cluster to build the seed, as explained in Section 5.1.2.
- **Source Classifiers' Entropy (*S*-H):** a source classifier is used to calculate the target instances' entropy score and select the ones with highest uncertainty as seed. Both classifiers trained using language dependent (S_D) and language independent (S_I) data are tested, using the different entropy scorers presented in Section 5.1.1 for each type.

Method	**B=100**	**B=250**	**B=500**	**B=1000**
Random	0.598 ±.024	0.735 ±.013	0.793 ±.004	0.829 ±.004
Length	0.749	0.811	0.834	0.851
K-MC	0.655 ±.016	0.8 ±.004	0.832 ±002	0.864 ±001
K-MC & Length	0.665 ±.021	0.794 ±.003	0.831 ±.004	0.864 ±.001
S_D-H Sum	0.746	0.809	0.838	0.845
S_D-H Mean	0.108	0.456	0.627	0.66
S_D-H K-Max	0.717	0.762	0.828	0.854
S_D-H Max	0.69	0.737	0.783	0.849
S_I-H Sum	0.777	0.806	0.831	0.862
S_I-H Mean	0.08	0.289	0.432	0.609
S_I-H K-Max	0.769	0.797	0.821	0.863
S_I-H Max	0.67	0.762	0.807	0.862
S_I-H Sum & K-MC	**0.79** ±.005	**0.839** ±.003	**0.858** ±.002	**0.876** ±.001
S_I-H K-Max & K-MC	0.756 ±.007	0.805 ±.005	0.845 ±.002	**0.879** ±.002
S_I-H Sum & Length	0.77	0.786	0.842	**0.878**
S_I-H K-Max & Length	0.766	0.786	0.842	0.873

Table 5: F1 and standard error results for the tested seed selection methods and sizes (B).

- ***S*-H and K-MC Combination:** the instances selected as seed are those in top positions when ranked according to their entropy and K-MC combination score using Equation 1 in Section 5.1.3.
- ***S*-H and Length Combination:** the instances are ranked according to their entropy and length combination score and the top ones are chosen as seed. The combination score of an instance is the product multiplication of its rescaled (range 0-1) length with its entropy score.

Table 5 shows the results obtained for each explored configuration using practical seed sizes for a real environment. For the K-Means and Random selectors, as they have a random component, their mean and standard error over 5 iterations are shown. As expected, the random baseline performs the worst. Selecting instances according to their length gives good results, although performance decreases as the number of selected instances in the seed increases. One possible explanation for this method's performance may be that longer utterances contain a higher number of tokens from which to learn, hence giving good results in small seeds, but simply querying the largest utterances does not account for data diversity nor content and tokens may repeat themselves in larger seeds, thus not providing new information due to content overlap.

The K-MC sampling method produces better results as the seed size increases, demonstrating that input diversity plays an important role for instance sampling. Nonetheless,

results for smaller seeds are lower than using other methods, since this methods does not take into account information content. Its combination with utterance length, which does not include content information either, does not improve the results obtained by the method in isolation.

When transferring knowledge from the source classifiers using the entropy scorers Sum and K-Max[4], S_I and S_D models perform quite similarly. S_I models perform slightly better in smaller seeds, although the difference gets narrower in larger seeds. While both S-H K-Max and S-H Sum prove to be useful, the latter is directly related to instance length, as the longer the sequence is the more likely it is to have a higher entropy sum. It also displays similar patterns to the utterance length method, no longer being among the top methods in the largest seed size tested. A possible explanation for this could be that it takes into account all the words of the instance, thus being sensitive to noise. On the other hand, the S-H K-Max scorer takes into account only the K words in the instance with highest entropy, so it is more agnostic to length and low-entropy words in the utterance, making it more robust to noisy instances. S-H Mean and S-H Max do not provide good results —especially the former, which produces results even worst than the weak random baseline—, and so are not good eligible seed selection methods. As S_I models yield slightly better results than S_D in general and only S-H K-Max and S-H Sum turned out to be eligible scorers, only those configurations were combined with length and K-MC for further explorations.

The best results for seed selection are obtained when combining K-MC with the S_I-H Sum scorer, as this method takes into account the divergence between the input data, the length of the input samples, and their uncertainty. Likewise, the combination of K-MC with the S_I-H K-Max model achieves good results in larger seed sizes. It is interesting to mention that when K-MC is combined with S-H the standard error intervals are reduced as compared with the K-MC method in isolation or in its combination with length, prominently in small seeds. This shows enhanced robustness of the method.

One real case scenario where the implications of choosing a good seed can be conceived is that of using automatic pre-annotation, i.e. making use of the predictions made by the current classifier to help the oracle in the query labelling process. The aim of pre-annotation is to help reduce the number of labelling actions required by the oracle, but if the classifier's predictions are wrong too often it can result in extra work for the oracle to correct the labelling errors. In the presented case, if the learner trained on the random seed is used for pre-labelling the number of corrections required for the oracle may be much greater than using one trained on a better seed, since in the case of the latter the performance of the learner is improved from the first training iteration.

5.4.2 Query Selection Strategy Evaluation

In this section, different query strategies are tested and compared in a simulated AL setting. In order to visualise the different methods' impact on training speed, the learning curves

[4]K=3 is used throughout all the experiments

for the base learners trained over the first 10k selected instances are plotted in Figure 3. The best configuration obtained in Section 5.4.1 was used as seed. The classifiers were asked to stop learning once they reached top performance[5]. Although they do not need to be the same, the selected size for both seed and query batches was 250 instances, which seems appropriate for a real case scenario.

The different methods evaluated are the following:

- **Random Query Selection:** passive learning, used as a weak baseline.
- ***T*-H Query Selection:** traditional uncertainty-based query strategy, using the base learner's uncertainty only. The two best sequence entropy scorers are used: H Sum and H K-Max.
- ***T*-H and *S*-H Combination Query Selection:** using the product multiplication of the source and target classifiers' uncertainty scores as query strategy. Since the two best entropy scorers are used, there are four possible combinations: (i) T-H Sum $\cdot$ S-H Sum, (ii) T-H Sum $\cdot$ S-H K-Max, (iii) T-H K-Max $\cdot$ S-H K-Max, and (iv) T-H K-Max $\cdot$ S-H Sum.
- **K-MC and *T*-H Combination Query Selection:** using the combination score of K-MC and the target classifier's entropy (following Equation 1) as query strategy. Again, the two best entropy scorers are used in the combination: H Sum and H K-Max.

As expected, the resulting learning curves presented in Figure 3 show that all the proposed active query selection criteria perform significantly better than passive random selection, exhibiting a much steeper curve and reaching top performance in fewer learning iterations.

The two traditional uncertainty-based methods which consider the target classifier's uncertainty only, i.e. T-H Sum and T-H K-Max, perform equally well and reach top performance using 4250 instances as training only. The top positions of best tested query strategies is occupied by models which combine both source and target classifiers' uncertainty. Concretely, combinations of both the T-H K-Max and the T-H Sum with S-H Sum reach the top performing score with just 4000 training instances in the labelled set, which constitutes the 12% of the corpus total training set. Therefore, both methods outperform traditional query selection strategies where only target classifier's uncertainty is taken into account. Using the T-H Sum $\cdot$ S-H K-Max combination performs just as well as traditional uncertainty sampling. On the other hand, the combination of source and target classifiers information which does not include the H Sum scorer, i.e. the T-H K-Max $\cdot$ S-H K-Max strategy, performs moderately worse than the mentioned methods, requiring extra iterations (one more than traditional methods and two more than the best tested methods) to reach top performance. Considering that H K-Max is agnostic to instance length, it may

[5]Instead of using the hard F1=0.935 top performance score, a breakpoint of F1=0.9345 was established to smooth the stopping criterion.

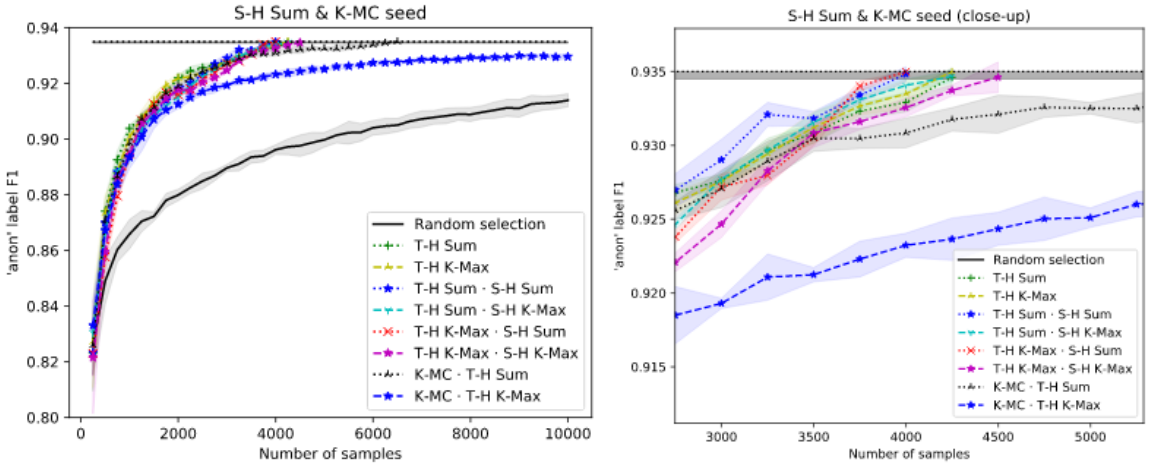

Figure 3: Learning curves using the tested query selection strategies for AL (left) and close-up look at the top performance reaching iterations of the best methods (right). Standard deviation over 5 iterations appears shadowed.

be concluded that this aspect could actually play a somewhat important role to achieve best query selection. Further support for this hypothesis is that the K-MC combination with T-H Sum produces much better results than its combination with T-H K-Max, which does not get to reach top performance even when using 10k training instances. These two methods perform the worst among the active query strategies tested, which suggests that K-MC introduces noise when used for query selection, even if it was among the best options for seed selection when combined with entropy.

The fact that those strategies which make use of the H Sum scorer generally get better results than those which do not may be simply due to them querying larger instances, i.e. utterances with a larger number of tokens. In the presented AL simulations an assumption is made that querying less instances implies less annotation time, effort and expenses. Although the aim of this work is to explore seed and query selection strategies and not cost evaluation, it would be interesting to check the effect of querying larger instances on annotation cost in a real AL setting as future work, for example considering the number of tokens in order to reduce or enlarge the size of the batches of the queried instances in each iteration.

6 Conclusions and Future Work

In this thesis, a new anonymised corpus was compiled from scratch and then exploited to investigate the automation of the anonymisation process in cases where little annotated data are available. As a result, new seed selection and query selection strategy methods combining Knowledge Transfer and Active Learning are proposed to approach annotated data scarcity for the textual anonymisation task.

The gathered corpus is a compilation of spontaneous spoken human-human dialogues in the telecommunications technical support domain in Spanish. The compilation process has been completed from the transcriptions of the raw audio files to their anonymisation (including 11 different anonymisation labels) and annotation of linguistic and extralinguistic phenomena (e.g. music, laughter, code switching, use of filler words, etc.). The compilation and anonymisation processes have been explained thoroughly step by step, including mentions to the difficulties found and considerations which could be useful for those who want to carry out similar tasks.

The ES-Port corpus is publicly available for research via META-SHARE, in the repository of the University of the Basque Country UPV/EHU[6] under the name *ES-PORT*. Notice that although the transcriptions have been properly anonymised, the raw audio corpus has not and therefore cannot be shared, as it contains sensitive data which fall under the GDPR.

The gathered corpus has been exploited in a sensitive data identification task, where the simulated scenario has included a small source annotated corpus from a different language and domain, ITAC, and a large target dataset, the ES-Port corpus, with no labels available. To accelerate the process of learning a robust classifier for the task in the unlabelled target domain, an Active Learning setting has been used. Uncertainty from a classifier trained on the source data has been employed both for seed selection and query selection strategy, using different sequence entropy scorers: Max Entropy, Mean Entropy, K-Max Entropy and Entropy Sum. Different methods considering uncertainty, input divergence, length, and their combinations have been tested and compared. For seed selection, methods have been tested using different seed sizes considered practical for a real case setting. Those methods which take all three aspects into account have obtained better seeds. For query selection strategy, uncertainty from the active base learner has been combined with that of the source classifier and compared with traditional uncertainty sampling (i.e. uncombined base learner's uncertainty) and with its combination with an input diversity method (K-Means Centroids). The proposed strategy combining source and target classifiers' uncertainties obtains the best results, reaching top classifier performance in fewer iterations using less than 12% of the full training set.

Although results of using the best methods combining source and target classifier's uncertainty reach top performance only one iteration (that is 250 instances in the case of the selected batch size) earlier than traditional uncertainty sampling, it must be noted that the source classifier has been trained on very little data (473 utterances only). In future

[6] `http://aholab.ehu.es/metashare/repository/search/`

work, it would be of interest to test whether using source models trained on more data can further improve query selection.

Future work regarding the ES-Port corpus could involve training NERC tools that work better on spontaneous dialogue data in Spanish or analysing self-correction and turn-taking strategies in a telephonic context. Annotating the corpus at dialogue act and other levels would also enrich the usefulness of the resource for dialogue system development and other purposes. With regard to the proposed approach related to exploiting classifier knowledge transfer for Active Learning, future work includes testing it in non-binary classification tasks. Particularly, it would be interesting to replicate the exposed experiments in a unified identification plus categorisation of sensitive data task, perhaps using ES-Port as target and ITAC as source again, even if such source does not contain labels regarding category. Other research options include exploring new ways to exploit multiple source corpora/classifiers from different domains, either for textual anonymisation or for other tasks.

References

Rodrigo Agerri, Josu Bermudez, and German Rigau. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *LREC*, volume 2014, pages 3823–3828, 2014.

Aitor Álvarez, Carlos-D Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. Improving the automatic segmentation of subtitles through conditional random field. *Speech Communication*, 88:83–95, 2017.

L Douglas Baker and Andrew Kachites McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103. ACM, 1998.

Jason Baldridge and Miles Osborne. Active learning and the total cost of annotation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1):5–22, 2001.

Eckhard Bick and Anabela Barreiro. Automatic anonymisation of a new Portuguese-English parallel corpus in the legal-financial domain. *Oslo Studies in Language*, 7(1), 2015.

Michael Bloodgood and K Vijay-Shanker. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 39–47. Association for Computational Linguistics, 2009.

Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 59–66, 2003.

Xavier Carreras, Isaac Chao, Lluis Padró, and Muntsa Padró. FreeLing: An Open-Source Suite of Language Analyzers. In *LREC*, pages 239–242, 2004.

Rita Chattopadhyay, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Joint transfer and batch-mode active learning. In *International Conference on Machine Learning*, pages 253–261, 2013.

Yukun Chen, Thomas A Lask, Qiaozhu Mei, Qingxia Chen, Sungrim Moon, Jingqi Wang, Ky Nguyen, Tolulola Dawodu, Trevor Cohen, Joshua C Denny, et al. An active learning-enabled annotation system for clinical named entity recognition. *BMC medical informatics and decision making*, 17(2):82, 2017.

Council of European Union. Article 29 Data Protection Working Party: Opinion 05/2014 on Anonymisation Techniques. 2014. URL http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

Council of European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), May 2016. URL http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf. Accessed July, 2018.

Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751, 2005.

Aron Culotta, Trausti Kristjansson, Andrew McCallum, and Paul Viola. Corrective feedback and persistent learning for information extraction. *Artificial Intelligence*, 170(14-15):1101–1122, 2006.

Francisco Dias, Nuno Mamede, and Jorge Baptista. Automated anonymization of text documents. In *IEEE World Congress Computational Computational Inteligence/Intelligence Methods for NLP*, pages 1287–1294, 2016.

Francisco Manuel Carvalho Dias. Multilingual Automated Text Anonymization. Master's thesis, Instituto Superior Técnico de Lisboa, 2016.

Dmitriy Dligach and Martha Palmer. Using language modeling to select useful annotation data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, pages 25–30. Association for Computational Linguistics, 2009.

Dmitriy Dligach and Martha Palmer. Good seed makes a good crop: accelerating active learning using language modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 6–10. Association for Computational Linguistics, 2011.

David A Dorr, WF Phillips, Shobha Phansalkar, Shannon A Sims, and John Franklin Hurdle. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of information in medicine*, 45(03):246–252, 2006. doi: 10.1055/s-0038-1634080.

Chunjiang Fu and Yupu Yang. Low density separation as a stopping criterion for active learning svm. *Intelligent Data Analysis*, 19(4):727–741, 2015.

Laura García-Sardiña, Manex Serras, and Arantza Del Pozo. ES-Port: a Spontaneous Spoken Human-Human Technical Support Corpus for Dialogue Research in Spanish.

In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018a. ISBN 979-10-95546-00-9.

Laura García-Sardiña, Manex Serras, and Arantza Del Pozo. Knowledge Transfer for Active Learning in Textual Anonymisation. In *International Conference on Statistical Language and Speech Processing (SLSP 2018)*, Mons, Belgium, October 2018b. Springer. In Press.

Masood Ghayoomi. Using variance as a stopping criterion for active learning of frame assignment. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 1–9. Association for Computational Linguistics, 2010.

Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. In *Advances in neural information processing systems*, pages 593–600, 2008.

Robbie Haertel, Eric Ringger, Kevin Seppi, James Carroll, and Peter McClanahan. Assessing the costs of sampling methods in active learning for annotation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 65–68. Association for Computational Linguistics, 2008.

Steven CH Hoi, Rong Jin, and Michael R Lyu. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web*, pages 633–642. ACM, 2006a.

Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424. ACM, 2006b.

Sheng-Jun Huang and Songcan Chen. Transfer learning with active queries from source domain. In *IJCAI*, pages 1592–1598, 2016.

Rebecca Hwa. Sample selection for statistical parsing. *Computational linguistics*, 30(3): 253–276, 2004.

Ashish Kapoor, Eric Horvitz, and Sumit Basu. Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning. In *IJCAI*, volume 7, pages 877–882, 2007.

Dimitrios Kokkinakis and Anders Thurin. Anonymisation of Swedish clinical data. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 237–241. Springer, 2007.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Florian Laws and Hinrich Schütze. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 465–472. Association for Computational Linguistics, 2008.

David Lewis and William Gale. Training text classifiers by uncertainty sampling. 1994.

Shoushan Li, Yunxia Xue, Zhongqing Wang, and Guodong Zhou. Active Learning for Cross-domain Sentiment Classification. In *IJCAI*, pages 2127–2133, 2013.

Xiao-Bai Li and Jialun Qin. Anonymizing and sharing medical text records. *Information Systems Research*, 28(2):332–352, 2017.

Harald Lüngen, Michael Beißwenger, Laura Herzberg, and Cathrin Pichler. *Anonymisation of the Dortmund Chat Corpus 2.1.* Institut für Deutsche Sprache, Bibliothek, 2017.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No.14*, pages 281–297. Oakland, CA, USA, 1967.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Ben Medlock. An introduction to NLP-based textual anonymisation. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), Genes, Italy*, 2006.

Hoang-Quoc Nguyen-Son, Minh-Triet Tran, Hiroshi Yoshiura, Noboru Sonehara, and Isao Echizen. Anonymizing personal text messages posted in online social networks and detecting disclosures of personal information. *IEICE TRANSACTIONS on Information and Systems*, 98(1):78–88, 2015.

Fredrik Olsson. *Bootstrapping named entity annotation by means of active machine learning: a method for creating corpora*. PhD thesis, 2008.

Fredrik Olsson. A literature survey of active machine learning in the context of natural language processing. 2009.

Fredrik Olsson and Katrin Tomanek. An intrinsic stopping criterion for committee-based active learning. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 138–146. Association for Computational Linguistics, 2009.

Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Rachel Panckhurst. A large sms corpus in french: From design and collation to anonymisation, transcoding and analysis. *Procedia-Social and Behavioral Sciences*, 95:96–104, 2013.

Namrata Patel, Pierre Accorsi, Diana Inkpen, Cédric Lopez, and Mathieu Roche. Approaches of anonymisation of an SMS corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 77–88. Springer, 2013.

Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32. Association for Computational Linguistics, 2010.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846. Citeseer, 2000.

Burr Settles. *Curious machines: Active learning with structured instances*. PhD thesis, University of Wisconsin–Madison, 2008.

Burr Settles. Active Learning Literature Survey. *Computer Sciences Technical Report*, 1648, 2010.

Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1070–1079. Association for Computational Linguistics, 2008.

Burr Settles, Mark Craven, and Lewis Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, pages 1–10. Vancouver, Canada, 2008a.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in neural information processing systems*, pages 1289–1296, 2008b.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

Claude E Shannon. A note on the concept of entropy. *Bell System Tech. J*, 27(3):379–423, 1948.

Hao Shao. Query by diverse committee in transfer active learning. *Frontiers of Computer Science*, pages 1–12, 2018.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 589. Association for Computational Linguistics, 2004.

Xiaoxiao Shi, Wei Fan, and Jiangtao Ren. Actively transfer domain knowledge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 342–357. Springer, 2008.

György Szarvas, Richárd Farkas, and Róbert Busa-Fekete. State-of-the-art anonymization of medical records using an iterative machine learning framework. *Journal of the American Medical Informatics Association*, 14(5):574–580, 2007.

Katrin Tomanek and Udo Hahn. Approximating Learning Curves for Active-Learning-Driven Annotation. In *LREC*, volume 8, pages 1319–1324, 2008.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. Efficient annotation with the jena annotation environment (jane). In *Proceedings of the Linguistic Annotation Workshop*, pages 9–16. Association for Computational Linguistics, 2007.

Katrin Tomanek, Florian Laws, Udo Hahn, and Hinrich Schütze. On proper unit selection in active learning: co-selection effects for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 9–17. Association for Computational Linguistics, 2009.

Amund Tveit, Ole Edsberg, TB Rost, Arild Faxvaag, O Nytro, T Nordgard, Martin Thorsen Ranang, and Anders Grimsmo. Anonymization of general practioner medical records. In *second HelsIT Conference*, 2004.

Andreas Vlachos. A stopping criterion for active learning. *Computer Speech & Language*, 22(3):295–312, 2008.

Zuobing Xu, Ram Akella, and Yi Zhang. Incorporating diversity and density in active learning for relevance feedback. In *European Conference on Information Retrieval*, pages 246–257. Springer, 2007.

Tong Zhang and F Oles. The value of unlabeled data for classification problems. In *Proceedings of the Seventeenth International Conference on Machine Learning,(Langley, P., ed.)*, pages 1191–1198. Citeseer, 2000.

Jingbo Zhu and Eduard Hovy. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. Learning a stopping criterion for active learning for word sense disambiguation and text classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008a.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. Multi-criteria-based strategy to stop active learning for data annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1129–1136. Association for Computational Linguistics, 2008b.

Jingbo Zhu, Huizhen Wang, Eduard Hovy, and Matthew Ma. Confidence-based stopping criteria for active learning for data annotation. *ACM Transactions on Speech and Language Processing (TSLP)*, 6(3):3, 2010.

Appendix I

The following table (6) contains the new set of Part of Speech (PoS) tags (*Normalised tag* column) and their correspondences for CoreNLP's output tags for Spanish (*Spanish tag* column) and English (*English tag* column). Information about the types of elements included in each PoS category is given in the final (*Information*) column.

Normalised tag	**Spanish tag**	**English tag**	**Information**
JJ	s.w. "a"	s.w. "JJ"	adjective
IN	s.w. "sp" or = "cs"	= "TO" / "IN"	preposition or subordinating conjunction
CC	= "cc"	= "CC"	coordinating conjunction
CD	s.w. "dn" or = "z0"	= "CD"	cardinal number
POS	s.w. "dp"	= "POS"	possessive determiner or ending
PUNC	s.w. "f" or in string.punctuation or = "word"	= "SYM"	punctuation or symbol
UH	= "i"	= "UH"	interjection
NN	s.w. "nc00" / "nc0n" / "nc0s" or = "w"	= "NN"	singular noun
NNS	s.w. "nc0p" or = "zu" / "zm"	= "NNS"	plural noun, measure
NNP	s.w. "np0"	= "NNP" / "NNPS"	proper noun
PRP$	s.w. "px"	= "PRP$" / "WP$"	possessive pronoun
PRP	s.w. "p"	= "PRP" / "WP"	pronoun
DT	s.w. "d"	= "DT"	determiner
RB	= "rg" / "rn"	s.w. "RB" or = "RWB"	adverb
VB	s.w. "v" or = "modal"	s.w. "VB"	verb
X	None of the above	None of the above	other

Table 6: Normalised Part of Speech tag correspondences for CoreNLP's Spanish and English outputs and information about the type of elements included in each. ["s.w.": *starts with*; "=": *is equal to*]

The correspondences are written as rules concerning the given tag, and they are displayed in the table in a hierarchical way, meaning the top rules have priority over the bottom rules. This way, if the given Spanish tag starts with "dn" its assigned tag will be "CD" and not "DT", since the rule *starts with "dn"* is placed above the rule *starts with "d"*.

The full original tag sets used by CoreNLP can be consulted in the following links:

- Spanish PoS tag set: `https://nlp.stanford.edu/software/spanish-faq.shtml#tagset`
- English PoS tag set: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

Appendix II

Next, the two publications related to the work presented in the current thesis are attached in their original format:

- **ES-Port: a Spontaneous Spoken Human-Human Technical Support Corpus for Dialogue Research in Spanish** (García-Sardiña et al., 2018a). Paper related to the ES-Port corpus compilation process, presented in the 11th edition of the Language Resources and Evaluation Conference (LREC) in May 2018.
- **Knowledge Transfer for Active Learning in Textual Anonymisation** (García-Sardiña et al., 2018b) [preview]. Paper related to the exploration of a new approach using Knowledge Transfer for Active Learning to speed up training supervised text anonymisation systems for large target unlabelled datasets. This paper will be presented in the 6th International Conference on Statistical Language and Speech Processing (SLSP) in October 2018. The attached version is just a manuscript. The final published version can be found in the cited publication.

ES-Port: a Spontaneous Spoken Human-Human Technical Support Corpus for Dialogue Research in Spanish

Laura García-Sardiña, Manex Serras, Arantza del Pozo
Speech and Natural Language Technologies, Vicomtech
Mikeletegi Pasealekua 57, 20009 Donostia-San Sebastián
{lgarcias, mserras, adelpozo}@vicomtech.org

Abstract

In this paper the ES-Port corpus is presented. ES-Port is a spontaneous spoken human-human dialogue corpus in Spanish that consists of 1170 dialogues from calls to the technical support department of a telecommunications provider. This paper describes its compilation process, from the transcription of the raw audio to the anonymisation of the sensitive data contained in the transcriptions. Because the anonymisation process was carried out through substitution by entities of the same type, coherence and readability are kept within the anonymised dialogues. In the resulting corpus, the replacements of the anonymised entities are labelled with their corresponding categories. In addition, the corpus is annotated with acoustic-related extralinguistic events such as background noise or laughter and linguistic phenomena such as false starts, use of filler words or code switching. The ES-Port corpus is now publicly available through the META-SHARE repository, with the main objective of promoting further research into more open domain data-driven dialogue systems in Spanish.

Keywords: spontaneous dialogue corpus, human-human dialogue, technical support, transcription, anonymisation, named entities

1. Introduction

Dialogue systems, often referred to as conversational agents or chatbots, are becoming increasingly popular as they allow users to directly interact with a wide range of information systems in a natural way. Customer support is an application scenario with a strong interest in dialogue system development, driven by the promise of intelligent digital assistants available 24x7 to resolve customer requests in a fast, cost-effective and consistent manner (Guzmán and Pathania, 2016).

Data-driven approaches to dialogue system development have shown to be more robust than rule-based techniques to variability in user behaviour, the performance of speech and language processing subcomponents and the dynamics of the task domain (Meena, 2015). Despite their promising results in recent years (Young et al., 2013; Wen et al., 2016; Li et al., 2017), most practical dialogue systems are still built by human experts through significant manual engineering. In most currently deployed systems, rule-based dialogue managers (DM) are combined with statistical natural language understanding (NLU) models capable of classifying intents and their related entities (Williams et al., 2015). In the customer support domain, this limits their application to frequent use cases in specific areas where solutions are well known, predictable and where scripted answers can be developed (Guzmán and Pathania, 2016).

Lack of annotated corpora is the main problem for the development of data-driven systems (Serban et al., 2015). To overcome this issue, it is usual to develop rule-based baselines or to employ the Wizard of Oz (WOZ) technique (Benedı et al., 2006; Rieser and Lemon, 2008) in which a human mimics the intended dialogue system, in order to gather interactions with real users. Although this kind of human-machine data is constrained by the employed baseline systems or the scenarios defined in the followed WOZ approaches, it is useful to bootstrap goal-driven dialogue systems whose policies can be optimised through user simulation and adaptive learning (Schatzmann et al., 2007; Gašić et al., 2013; Serras et al., 2017). On the other hand, human-human dialogue corpora contain unconstrained and unscripted natural dialogue interactions exhibiting traits different from human-machine dialogue (i.e. richer turn-taking and more common grounding phenomena) (Doran et al., 2003), which are more suitable to train more open domain dialogue systems. Human-human customer support corpora would allow progress towards the development of large-scale data-driven dialogue systems capable of handling a wider amount of customer queries.

A considerable amount of corpora are available for building data-driven dialogue systems (Serban et al., 2015). Unfortunately, because customer support interactions occur in commercial settings, most customer support datasets are proprietary and not released to the public due to privacy and data protection reasons. In practice, there are only a couple of publicly available technical support datasets (Lowe et al., 2015; Uthus and Aha, 2013) derived from the Ubuntu IRC channels[1], used to receive technical support for issues related to the Linux-based operating system. Although some data is available from the support channels in other languages, most of the compiled resources are in English.

In this paper, the Spanish Technical Support (ES-Port) Corpus is presented, a compilation of spontaneous spoken human-human dialogues from the technical customer support service of a Spanish telecom operator for companies. The corpus has been directly transcribed from call recordings, annotated at various linguistic and acoustic-related extralinguistic levels, and anonymised in order to comply with data protection legislation. Its release is intended to foster further research into more open domain data-driven di-

[1]These logs are available from 2004 to 2018 at http: //irclogs.ubuntu.com/

alogue systems in Spanish, capable of achieving more natural interactions in the technical support domain.

2. Compilation Process

The raw corpus was provided by an independent telecom operator, dedicated to providing tailor made cloud data centre, fixed voice, IP or mobile telephony and Internet connectivity solutions to companies. In order to serve their clients, they offer 24/7 customer support: 24 hours a day, 7 days a week, 365 days a year. Despite having a multichannel customer service and also providing support through web forms and email, the majority of the clients still prefer calling. Thus, the corpus provided consisted of raw audio recordings of such calls.

2.1. Transcribing the Audio

The first step of the corpus compilation process involved transcribing the provided raw audio data. Details regarding the characteristics of the audio and the followed transcription process are given next.

2.1.1. Audio characteristics

The recorded calls contain both speech and other background sounds, such as channel-associated noises or background music. The type of speech used is spontaneous, and so it includes phenomena such as false starts, mispronunciations, non standard forms, overlapping segments between speakers, unfinished sentences and, in general, speech more focused on conveying the message than on taking care of its form. The audio data consisted of a total of 40 hours , with an average length of 2 minutes per dialogue.

2.1.2. Transcription process

The provided recordings were transcribed using the Transcriber 1.5 annotation tool (Barras et al., 2001). In addition to the orthographic transcriptions, the following phenomena were also annotated:

- speaker turns
- non-speech events (e.g. coughing and laughter) and background acoustic conditions (e.g. noise and music)
- overlapping speech
- false starts, repetitions, unfinished words and non-words
- mispronunciations, lengthening in pronunciation, and typical spoken Spanish shortening of words (e.g. *pa* instead of *para*) or dropping of intervocalic *d* in final syllables (e.g. *demasiao** for *demasiado*, *entrao** for *entrado*)
- continuers and filler words (e.g. *o sea*, *eh*, *hala*, *mhm*, etc.)
- words in a language other than Spanish (when pronounced correctly)

Given the more challenging spontaneous and telephone nature of the data, attempts to follow incremental automation methodologies such as those described in (Pozo et al., 2014) to make the transcription process more productive were not feasible. The word error rates (WER) of generic large vocabulary continuous speech recognition (LVCSR) systems turned out too high to provide any time savings (77.21% in test set).

In the end, the transcription process was carried out fully manually and took a linguist six months working full-time to complete.

2.2. Anonymising the Dialogues

In order to comply with the European data protection legislation (Art29WP, 2014) and not to compromise the right to confidentiality of the individuals involved, the personal information contained in any dataset must be neutralised before releasing the data open to the public in order to be exploited for other purposes.

Data anonymisation is the process of treating personal data in such a way that it can no longer be used to identify the individuals involved, while preserving the value and usefulness of the original format.

2.2.1. Anonymisation practices and standards

Despite European legislation does not prescribe any particular anonymisation technique, randomisation and generalisation approaches are usually employed to anonymise structured datasets in the form of tables or graphs:

- Randomization: involves alteration of the data without losing its value and includes techniques such as noise addition and permutation.
- Generalisation: implies diluting or reducing the granularity of the data and comprises techniques such as aggregation and K-anonymity.

For unstructured text, such as the transcriptions of the technical support recordings in the ES-Port corpus, the following methods have also been proposed in (Dias, 2016):

- Suppression: the element to be anonymised is replaced by some neutral indicator, e.g. 'XXXXX'.
- Tagging: the element to be anonymised is replaced by a label which can refer to its class or identifier, e.g. 'ORGANISATION123'.
- Substitution: the entity to be anonymised is substituted by another entity, e.g. 'Juan' for 'Pedro'. The choice of the new entity can be random from a dictionary, swapped with another entity within the document, 'intelligently' substituted by an entity sharing the same features, or applying a generalisation technique to the item (e.g. replacing 'University of the Basque Country' by 'University').

The technique chosen to anonymise the ES-Port corpus was substitution because readability and coherence are kept and the result is a natural anonymised text. The process is described in Section 2.2.3..

Table 1: NER and NERC Precision (Pr), Recall (Rc), and F1 scores for the three taggers on our test set.

	NER			NERC		
	Pr	Rc	F1	Pr	Rc	F1
IXA Pipes	0.32	0.62	0.42	0.26	0.50	0.34
FreeLing	0.36	0.65	0.47	0.24	0.54	0.33
CoreNLP	0.47	0.96	0.63	0.36	0.99	0.53

2.2.2. Identifying the features to be anonymised

The first step in the anonymisation process is to identify the type of elements in the dataset that refer to personal information or that could possibly be used in any way to identify the people involved, endangering their right to confidentiality. Considering the nature of the information given in the ES-Port corpus, we decided to anonymise the following types of elements:

- Elements referring to individuals' basic personal information: names, surnames, name diminutives or nicknames, personal identification numbers.
- Contact information and digital trace elements: phone numbers[2], IP addresses, user names and numbers, email addresses, postal addresses, web domains.
- Workplace and organisation-related elements: names of organisations, NIFs (tax identification number) and CIFs (tax code), easily linkable names of products and services, prices.
- Other elements: card numbers and bank accounts, dates, locations, trouble ticket numbers, dispatch notes, passwords, spellings of any of the previous elements.

2.2.3. Anonymisation process

Once the types of elements that needed to be anonymised were identified, the anonymisation process was carried out in a semi-automatic way.

First, the items to be anonymised were selected and categorised. We tried to automate the selection process by using different Named Entity Recognition and Classification (NERC) tools available for the Spanish language, namely IXA Pipes (Agerri et al., 2014), FreeLing (Carreras et al., 2004) and Stanford CoreNLP (Manning et al., 2014). Although these taggers have reported good results on planned written language such as news texts, trial tests on a small dataset of our spontaneous spoken technical support corpus were too poor to automate the process of selecting and categorising the items to be anonymised, as their use would still require considerable manual revision and correction. Results for the three taggers on the test set both considering entity recognition alone (NER) and entity classification as well (NERC) are presented in Table 1. In addition, none of the taggers covered all types of items that had to be anonymised, as is the case of numbers, months and individual letters in spellings.

The next step was automatic and consisted in randomly substituting the selected items for an element of the same characteristics according to its type (organisation, number, male/female name, etc.). Once an item was anonymised, its substitution was kept throughout the whole dialogue in order to maintain coherence, but not across dialogues so as to prevent possible linkability issues. New names provided for organisations and domains are made up and did not correspond to any existing entity at the time the anonymisation was carried out. The final step involved manual revision of the results and correction of coherence errors (e.g. non matching spellings).

[2]Some prefixes were kept if relevant to the conversation.

Table 2: Entity tags used in the anonymisation process.

Utterance	**Replacements**	**Tags**
"Soy Bárbara de Cincode"	Bárbara	female_name
	Cincode	organisation
"Arturo Noriega arroba Hotmail punto es, tengo que poner?"	Arturo	male_name
	Noriega	surname
	Hotmail	mail
"te la digo, es M de Madrid,"	M	letter
	Madrid	place
"el último registro es del veintisiete de septiembre."	veintisiete	number
	septiembre	month
"Inexistent punto com."	Inexistent	domain
"tiene que entrar= a CompDNS"	CompDNS	product/ service

The named entity categories of the elements anonymised following the process described above have been kept in the compiled corpus. As a result, ES-Port also includes named entity annotations. However, these were anonymisation-oriented and therefore the number of classes and specificity of the tags are more granular than in the typical NERC approaches. Nevertheless, the tags used are easily generalisable to the usual four (PERSON, LOCATION, ORGANISATION, MISCELANEA) or six (plus NUMBER and DATE) NERC classes. The tags used and real examples of their usage are shown in Table 2.

Overall, the described anonymisation process took a linguist eight months working part-time to complete.

3. The Compiled Corpus

This section describes the gathered corpus quantitatively and qualitatively.

3.1. The corpus in numbers

Table 3 summarizes the basic statistics of the ES-Port corpus regarding its number of dialogues, turns and overlaps, its vocabulary and its amount of filler and foreign words.

Table 3: Corpus Characteristics

Num. Dialogues	1170	Vocabulary size	11221
Num. Turns	65239	Labelled Filler Words	37
Avg. Turns per Dialogue	55.76	Filler Words Freq.	26574
Avg. Turn Length	8.20	Foreign Words Freq.	3294
Num. Overlap Turns	11329	English Words Freq.	3017

As can be seen, the corpus presents attributes typical of spontaneous spoken human-human interaction such as overlapping turns (around 17% of the turns) and rich use of

Table 4: Excerpt from a dialogue, including turn index (T), speakers (S) and the actual annotated and anonymised utterance transcription (U).

T	S	U
29	spk1	De todas formas
	spk2	esto
30	spk1	si has +enviado el correo estate tranquilo porque <se=> se para.
31	spk2	(*EVENT*: noise-rire)
	spk2	<%mm> Es <lom-> <i-> incluso si lo <envi-> <%aver> <su-> supuestamente hasta las cinco y media, no?
32	spk1	Sí.
33	spk2	(*EVENT*: noise-rire)
	spk2	Y si lo envío a las cinco y diez se cancela?
34	spk1	Sí, sí.
35	spk2	Ay, dios (*EVENT*: pronounce-ch)
36	spk1	<%mhm>
37	spk2	Ay, mi madre (*EVENT*: pronounce-ch)
	spk2	no puedo largarme de aquí digamos.
38	spk1	<%eh> si quieres
	spk2	<%eh>
39	spk1	llamar un poquillo más tarde y te intento pasar con él de nuevo.
40	spk2	Es que no hay ninguna forma de que, ningún número que yo pueda=
41	spk1	No, no.
	spk2	llamarles
42	spk2	a ellos o
43	spk1	No, <%osea>
	spk2	algo?
44	spk1	que le estoy llamando yo y no me responde.
45	spk2	(*EVENT*: noise-nontrans)
	spk2	<ueh->
	spk2	okay (*LANGUAGE*: en)
	spk2	<%pues> muchas gracias.
46	spk1	<%venga> a ti.

continuers and filler words (approximately 5% of total word occurrences). It is interesting to note that since the corpus is gathered from an IT domain, English foreign words are quite common, reaching up to 91.59% of all foreign words occurrences and constituting around 3.24% of the vocabulary.

3.2. Sample data description

In order to give an idea of the type of information and phenomena present in the ES-Port corpus, Table 4 shows an excerpt of one of its dialogues.

Different types of speech and non-speech events and background acoustic conditions occur often along the corpus. We find instances in turns 35 and 37, where the event tags indicate that the utterance was whispered, and in turn 45, indicating that the speech was unintelligible and could not be transcribed. Other instances can be found in turns 31 and 33, where the speaker laughs before continuing talking.

Overlapping speech is also common. Examples can be seen in turns 29, 41, and 43, where two different speakers intervene within the same turn. Speech, extralinguistic phenomena, or a combination of both can be overlapped.

False starts, repetitions, incomplete words, and nonwords are very common. False starts and repetitions are tagged simply using `<word>`, as in turn 30, while incomplete words and other nonwords are tagged as `<nonword->` and can be found in turns 31 and 45 of the excerpt.

Deviations from the standard pronunciation take place regularly in the corpus. An instance dropping the intervocalic phoneme /d/ in a final syllable can be found in turn 30, marked with a + symbol. Lengthening examples appear in turns 30 and 40, marked with an = symbol.

The words tagged following the pattern `<%word>` correspond to a set of 37 filler words (e.g. *o sea*, 'I mean') and continuers (e.g. *mhm*) frequently used in the corpus. Some instances of these can be found in turns 31, 36, 38, 43, 45, and 46.

Finally, words from a language other than Spanish appear quite often, especially in English. The language event in turn 45 indicates that a foreign word was used, in this case from English.

3.3. Potential Applications

The ES-Port corpus is a source of annotated spontaneous spoken human-human dialogues which may be valuable for several research tasks and applications. In this section, a few of them are mentioned.

Our main objective is to promote more open and natural dialogue interactions in Spanish customer support. Although it does not yet include dialogue act annotations, the corpus as is could be used to explore unsupervised approaches to dialogue system development in Spanish. These approaches include modelling the language of the system to generate more human-like prompts, modelling the language of the user to better detect the nuances of human-human communication or analysing the turn-taking dynamics for incremental dialogue processing.

The corpus annotated at the current level can also be exploited to develop supervised approaches for spontaneous LVCSR in Spanish (including automatic capitalization and punctuation) or to develop NERC tools that work better on dialogue text.

Finally, linguistic research in Spanish may use this data to study a wide range of issues, such as the use of discourse markers and filler words in conversation, their meaning in context, and how they influence the dialogue, or the strategies used for turn-taking and self-correction, among others. Other interesting phenomena for study are code switching and the abundant use of words from the English language in the IT domain.

3.4. Data sharing

The current version of the ES-Port dialogue corpus is available via META-SHARE[3] in the repository of the University

[3]The raw audio corpus cannot be released for public access, since it contains sensitive data which falls under the European General Data Protection Regulation (GDPR)

of the Basque Country UPV/EHU[4] under the name of ES-PORT.

4. Conclusions and Future Work

A spontaneous spoken human-human technical support dialogue corpus in Spanish has been transcribed and anonymised. At this point, the corpus contains annotations referring to linguistic and acoustic-related extralinguistic phenomena such as music, laughter, use of filler words and code switching in conversation. Named entities anonymised using the substitution technique are also annotated. The ES-Port corpus is now publicly released so it can be used for dialogue research or the adaptation of LVCSR and NERC systems to spontaneous dialogue. Future work includes dialogue act annotation of the corpus.

5. Acknowledgements

The authors would like to thank the telecom operator who has kindly provided the raw technical support call recordings. We would also like to thank Aholab, the Signal Processing Laboratory of the University of the Basque Country, for granting us access to their META-SHARE node to share our corpus.

6. Bibliographical References

Agerri, R., Bermudez, J., and Rigau, G. (2014). Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *LREC*, volume 2014, pages 3823–3828.

Art29WP. (2014). Article 29 data protection working party: Opinion 05/2014 on anonymisation techniques.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1):5–22.

Benedı, J.-M., Lleida, E., Varona, A., Castro, M.-J., Galiano, I., Justo, R., López, I., and Miguel, A. (2006). Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1636–1639.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *LREC*, pages 239–242.

Dias, F. M. C. (2016). Multilingual automated text anonymization.

Doran, C., Aberdeen, J., Damianos, L., and Hirschman, L., (2003). *Comparing Several Aspects of Human-Computer and Human-Human Dialogues*, pages 133–159. Springer Netherlands, Dordrecht.

Gašić, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P., and Young, S. (2013). On-line policy optimisation of bayesian spoken dialogue systems via human interaction. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8367–8371. IEEE.

Guzmán, I. and Pathania, A. (2016). Chatbots in customer service. *Accenture Interactive*.

Li, X., Chen, Y.-N., Li, L., and Gao, J. (2017). End-to-End Task-Completion Neural Dialogue Systems. *ArXiv e-prints*, March.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, abs/1506.08909.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Meena, R. (2015). *Data-driven Methods for Spoken Dialogue Systems*. Ph.D. thesis, KTH, Royal Institute of Thechnology.

Pozo, A. D., Aliprandi, C., Álvarez, A., Mendes, C., Neto, J. P., Paulo, S., Piccinini, N., and Raffaelli, M. (2014). Savas: Collecting, annotating and sharing audiovisual language resources for automatic subtitling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May.

Rieser, V. and Lemon, O. (2008). Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation. In *ACL*, pages 638–646.

Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., and Young, S. (2007). Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152. Association for Computational Linguistics.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. *CoRR*, abs/1512.05742.

Serras, M., Torres, M. I., and Del Pozo, A., (2017). *Online Learning of Attributed Bi-Automata for Dialogue Management in Spoken Dialogue Systems*, pages 22–31. Springer International Publishing, Cham.

Uthus, D. and Aha, D. (2013). The ubuntu chat corpus for multiparticipant chat analysis.

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2016). A Network-based End-to-End Trainable Task-oriented Dialogue System. *ArXiv e-prints*, April.

Williams, J., Kamal, E., Ashour, M., Amr, H., Miller, J., and Zweig, G. (2015). Fast and easy language understanding for dialog systems with microsoft language understanding intelligent service (luis). In *Proceedings of 2015 SIGDIAL Conference, Prague*. ACL – Association for Computational Linguistics, September.

Young, S., Gasic, M., Thomson, B., and Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems: A review. In *Proceedings of the IEEE*, volume 101(5), pages 1160–1179.

[4] http://aholab.ehu.es/metashare/repository/search/

Knowledge Transfer for Active Learning in Textual Anonymisation

Laura García-Sardiña, Manex Serras, and Arantza del Pozo

Speech and Natural Language Technologies, Vicomtech
Mikeletegi Pasealekua 57, 20009 Donostia-San Sebastián
{lgarcias, mserras, adelpozo}@vicomtech.org

Abstract. Data privacy compliance has gained a lot of attention over the last years. The automation of the de-identification process is a challenging task that often requires annotating in-domain data from scratch, as there is usually a lack of annotated resources for such scenarios. In this work, knowledge from a classifier learnt from a source annotated dataset is transferred to speed up the process of training a binary personal data identification classifier in a pool-based Active Learning context, for a new initially unlabelled target dataset which differs in language and domain. To this end, knowledge from the source classifier is used for seed selection and uncertainty based query selection strategies. Through the experimentation phase, multiple entropy-based criteria and input diversity measures are combined. Results show a significant improvement of the anonymisation label from the first batch, speeding up the classifier's learning curve in the target domain and reaching top performance with less than 10% of the total training data, thus demonstrating the usefulness of the proposed approach even when the anonymisation domains diverge significantly.

Keywords: Knowledge Transfer · Active Learning · Seed Selection · Query Selection Strategy · Textual Anonymisation

1 Introduction

Due to the growing amount of data (and especially textual data) created every day through social network posts, official documents, etc. that contain personal information, data privacy has gained a lot of attention over the last few years. Furthermore, valuable data which could be beneficial for research or transparency purposes may be kept unshared if it contains personal information because of the prohibitively high costs of its manual anonymisation and the legal repercussions of not doing it correctly. Even if datasets are not too large, manual anonymisation is a tedious and time-consuming task: Dorr et al. [6] assessed that manually de-identifying medical notes containing an average of 7.9 Personal Health Information items took around 87.3 seconds per note to complete. In this scenario, the automation of data sanitisation while preserving its usefulness has been widely researched [28, 12, 4, 8].

Different approaches oriented to the anonymisation of unstructured textual data have been proposed in [16, 4], where techniques of suppression, tagging/categorisation, and substitution are described. In this paper, the step previous to applying these techniques, i.e. personal data identification, is tackled. This step can be seen as a binary classification task, where the positive label corresponds to the words in an utterance that refer to sensible data such as personal names, organisations, passwords, and so on that need to be anonymised.

Lack of annotated data is a common issue when automating de-identification in supervised machine learning (ML) settings. In this context, the use of Active Learning (AL) [3] can optimise the data annotation phase, resulting in better ML models with fewer data. In a typical pool-based AL scenario the input is a small set of labelled instances (*seed*) and a large set of unannotated ones (*pool*). A classifier (*base learner*) is trained on the labelled instances and then asks an *oracle* to label the instance (in *serial* AL) or set of instances (in *batch mode* AL) which the classifier considers more informative according to some criterion (*query selection strategy*). The newly labelled data are moved from the pool to the labelled set and the classifier is retrained following this process iteratively until some *stopping criterion* is satisfied or the pool is empty. Two of the main questions that need to be answered in every AL framework are: (1) how to select the seed, and (2) which AL query selection strategy will be best to speed up the classifier's learning curve.

Very little attention has been paid to the seed selection aspect in the literature. Olsson [17] compared using a random seed against using cluster-centroid based sampling with little to no improvement for a NERC annotation task. Tomanek et al. [30, 29] compare multiple kinds of seeds checking instances against manually created entity gazetteers, reporting significant improvements over the random selection. Other automatic approach presented by Dligach and Palmer [5] uses unsupervised language model (LM) sampling to select a seed containing the examples with lowest LM probability in a word sense disambiguation task, obtaining significantly better results than using a random seed.

As surveyed by Settles [21, 22], there are multiple approaches for serial query selection in AL. In Uncertainty Sampling [10] scenarios, the learner uses an uncertainty measure (e.g. entropy) to query the most uncertain instances. Query-by-Committee [24] strategies use a committee of classifiers that present different hypotheses and query the instances with most disagreement. Expected classifier change methods query those instances which may cause the greatest change in the classifier. In Expected error reduction methods the classifier estimates the expected future error of the instances in the unlabelled pool and queries those with the minimal expected risk. In variance reduction strategies the learner queries those instances which minimise the output variance and thus the classifier's generalisation error. Finally, Density-weighted methods [2, 23, 20] query those instances which are both uncertain to the classifier and representative of the data's underlying distribution. Batch mode query strategies also attempt at selecting the best batch taking into account notions like information overlap in the set. Not much attention has been paid to this type of strategies in the

literature even if batch mode AL is a more realistic practice scenario, as the overhead of re-training the ML model for each annotated instance often renders serial query selection unusable.

The Knowledge Transfer (KT) or Transfer Learning paradigm encompasses the idea of re-using existing annotated resources to improve learning in new domains or tasks [18]. As the anonymisation task may re-use information extracted from different corpora that may vary in domain and language, it is sensible to consider combining KT with AL. There have been some previous works in the Natural Language Processing (NLP) field that combine KT with AL. Rai et al. [19] propose hyperplane-based distances to choose the most divergent samples from the source and target domains as seed in a sentiment analysis task. However the existence of this hyperplane narrows down the possibilities of classification algorithms and may not be suitable for sequential data [1]. Shi et al. [27] use a set of labelled instances in the target domain to train a text classifier with data from both domains, the oracle is only asked to label when the classifier's confidence is too low. In a sentiment classification task, Li et al. [11] train one classifier on the source data and another one on the target data and then both are used to select the most informative samples using a Query-By-Committee strategy.

The motivation of this paper is to speed up the process of training a robust classifier for textual data anonymisation using KT from available corpora within the Active Learning framework. Our main contribution is a previously unexplored method for transferring the knowledge from a classifier trained on a source corpus, differing from the target corpus both in language and domain, to improve the AL process both at seed selection and query selection strategies, and accelerating the learning curve in the target domain from the very first labelled batch. The source classifier's uncertainty is combined using different scoring methodologies to select the best possible seed and query selection criteria. Also, to the extent of our knowledge, the Active Learning paradigm is tested for the first time in an anonymisation task. Finally, a strong baseline for the anonymisation task using the publicly available ES-Port corpus [7] is set.

The paper is structured as follows: in Section 2 the proposed methods to exploit Knowledge Transfer for Active Learning from a theoretical point of view are described; then the feature sets and corpora used for the selected anonymisation task are introduced in Section 3; in Section 4 the methods are tested and their results are presented focusing on the two topics of interest of the paper: seed and query selection strategy in the AL setting; final remarks and conclusions are given in Section 5, as well as some ideas on future work directions.

2 Knowledge Transfer for Seed and Query Selection Strategy

In this section, the different query strategies used in this work and how the knowledge from the source domain is used to improve the Active Learning process are explained in detail.

2.1 Active Learning

The traditional pool-based Active Learning process as described in Section 1 is shown in Algorithm 1.

Algorithm 1: Pool Active Learning typical setting

input : set of labelled instances L, pool of unlabelled instances U, query strategy ϕ, batch size B, stopping criterion S

repeat

  // Train model M on L

  Q = best set in U of size=B according to ϕ

  // Ask Oracle to label Q

  $L = L + Q$

  $U = U$ - Q

until *S or size(U)=0*

return M, L

The anonymisation task is approached as a binary classification problem, where the positive label corresponds to the words to anonymise. Due to the sequential nature of the task, a discriminative model based on Conditional Random Fields [9] is used. These models have been intensively used for sequence labelling and segmentation [25, 15, 1].

2.2 Entropy Score Query Strategies

Being $I = (w_1, w_2, ...w_{|I|})$ an instance (i.e., a sentence or utterance) composed of words of a corpus and given the stochastic nature of the CRF classifiers, the uncertainty over the binary decision for each word $w_i \in I$ can be measured using the Shannon entropy [26]:

$$H(w_i) = -P(\hat{y}_i = A|\ I)log_2(P(\hat{y}_i = A|\ I))$$

$$-(1 - P(\hat{y}_i = A|\ I))log_2(1 - P(\hat{y}_i = A|\ I))$$

where $P(\hat{y}_i = A \mid I)$ is the probability of the classifier assigning the *anon* label A to the word w_i. As each instance is a sequence of words, the entropy score of the whole instance can be defined in multiple ways:

1. **H Sum:** Sum of all its word entropies: $H(I) = \sum_{w \in I} H(w)$
2. **H Mean:** Mean of its word entropies: $H(I) = \frac{1}{|I|} \sum_{w \in I} H(w)$
3. **H K-Max:** Mean of its K-Max word entropies: $H(I) = \frac{1}{K} \sum_{i=0}^{K} H(w)$, where the K words with highest entropy of the instance I are chosen.
4. **H Max:** Maximum entropy: $H(I) = max_{w \in I} H(w)$

The entropy scorers can be used to measure how certain the classifier is about a taken decision, yielding a robust query strategy to select the instances with high information content in the AL process.

2.3 K-Means-Centroids Query Strategy

The K-Means clustering algorithm [13] can be used to split the sample set into K clusters or groups. Then, the closest candidate to each cluster's centroid is selected. Being B the batch size of the instances to select from the pool, let $K = B$ in the clustering algorithm, splitting the pool in B clusters. Then, being $c_1, c_2, \cdots, c_B$ the centroids of each cluster and I_{c_k} the instances that encompass the cluster of centroid c_k, the closest instance I_k to the cluster centroid according to the Euclidean distance is chosen for each cluster:

$$I_k = argmin_{I \in I_{c_k}} ||c_k - I|| \; \forall \; k = 1, \cdots, B$$

2.4 K-Means-Centroids-Entropy Query Strategies

As the K-Means algorithm measures the input diversity and the $H(I)$ entropy scorers measure the base learner's uncertainty, both measures can be combined to select the instance I_k for each cluster centroid:

$$I_k = argmin_{I \in I_{c_k}} ||c_k - I|| \cdot (1 - rescale(H(I))) \tag{1}$$

where the $H(I)$ results are rescaled so they are within the range [0, 1].

2.5 Entropy-based Knowledge Transfer

In this section the proposed Knowledge Transfer methodology is explained. As depicted in Fig. 1, the entropy measures from the source classifier (S-H Sum/Mean/K-Max/Max) are used for both seed selection and query strategy in the target domain.

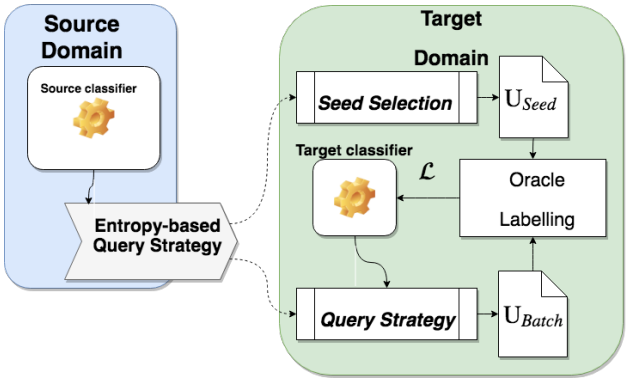

Fig. 1. KT schema, where entropy-based query strategies from the source classifier are used first for seed selection and then for query selection in the target domain

For seed selection, one cannot rely on knowledge from the target classifier or base learner since there is no labelled data in the target domain on which to train it. To overcome this limitation, the source domain classifier's entropy score S-H can be used to sample the most uncertain instances ($\mathcal{U}_{seed}$) of the target domain as seed. After annotation, these labelled instances $\mathcal{L}(\mathcal{U}_{seed})$ can be used to start training the target base learner. In addition, S-H can be combined with the target domain classifiers' entropy scorers (T-H) for query selection, in order to select the next batch $\mathcal{U}_{batch}$ of instances for the oracle to label.

3 Corpora

Two publicly available anonymised corpora differing in language, style, and domain have been used: ITAC and ES-Port. Both resources are briefly described below and their main characteristics are summarised in Table 1.

Table 1. ITAC and ES-Port corpora comparison

Characteristics	**ITAC**	**ES-Port**
Main Language	English	Spanish
Language Switching	No	Yes
Language Form	Written, planned (emails)	Spoken, spontaneous (phone calls)
Domain	Various (personal, corporate)	IT, Telecommunications
Training Utterances	473	47073

3.1 ITAC

The Informal Text Anonymisation Corpus (ITAC) [16] consists of about 2500 personal emails written in English. Due to the nature of the data, spelling, punctuation, and capitalisation inconsistencies and errors are common.

The corpus is anonymised with binary labels (anon/no-anon) and partitioned into training, development, and test sets of 666138, 6026, and 31926 tokens respectively. Unfortunately, only the last two sets are annotated. Following the solution given in [16] to the unannotated training set issue, the development set is used as training set.

Given the subjectivity of what constitutes a sensitive item that needs to be anonymised, ITAC was annotated following two different schemes: a comprehensive one where every reference that might possibly be related to people or organisations is anonymised even if the risk of identification is very low, called *blanket* anonymisation, and a more selective one where only those references directly related to people or organisations are annotated, referred to as *selective* anonymisation. In this work the blanket version is used as source corpus.

3.2 ES-Port

The Spanish Technical Support (ES-Port) corpus [7] consists of transcriptions of 1170 dialogues from calls to the technical support service of a telecommunications provider. Due to its nature, the corpus includes numerous turn overlaps, unfinished sentences and words, mispronunciations, filler words, grammatical errors, and other phenomena alike typical of spontaneous spoken language. Although Spanish is the main language of the corpus, various code switching events take place adding up to six other languages, of which English is the most common one.

The corpus is fully anonymised by token substitution. The types of items which are anonymised include basic personal information, contact information and digital trace items. Despite the anonymised items are annotated with their specific anonymisation categories, for the experiments reported in this paper the categorised labels have been converted to a simple 'anon' label to accommodate to our binary identification task.

As opposed to ITAC, ES-Port is not pre-partitioned, so for our tests we chose to divide the corpus by taking the first 900 dialogues (47073 utterances after the removal of turns not containing any text, e.g. silences, unintelligible speech) as training set and the rest (around 23% of the data) as test set.

3.3 Feature Selection

As the proposed methodology is used on cross-lingual data, two source classifiers were trained over the ITAC corpus, one with language independent features and another one with language dependent features. Beginning/End of Sentence (BOS/EOS), punctuation, case, NERC and Part of Speech (PoS) tags[1] were used as language independent features. For the language dependent case, features also included lower cased word forms and prefixes and suffixes (two and three first and last characters in the word). The selected features in a [-2, +2] word context window were also included. The features used for each instance (i.e., sentence or utterance) are the concatenation of the word-level features for each token in the sequence. The target classifier was trained over the ES-Port corpus with language dependent features only.

4 Experiments

In this section, the experiments carried out and their results are presented. Since the ITAC annotated training set is too short (473 utterances), we have tested the KT for AL setting using ES-Port as target, but not in the opposite direction. That being so, we will be referring to ITAC as *Source* and to ES-Port as *Target*. All results are reported on the ES-Port test set, taking into account the positive ('anon') label only.

[1] NERC and PoS tags were automatically extracted using the Stanford CoreNLP tool [14] for both languages and normalised to share the same values, e.g. both Spanish tag 'LUG' and English tag 'LOCATION' refer to place entities.

CRF classifiers were trained passively on the whole ES-Port training data to test their top performance, achieving 0.935 of F1 score on the 'anon' label. Source CRF models were trained using the blanket data, achieving 0.803 of F1 score with language dependent features and 0.785 without on ITAC's test set.

4.1 Seed Selection Evaluation

For seed selection, various methods have been implemented:

- **Random:** the seed is selected at random. This is used as a weak baseline.
- **Maximum Utterance Length:** the samples with largest number of words are chosen as seed.
- **K-Means-Centroids (K-MC):** the K-Means algorithm is used to split the corpus and choose a representative sample in each cluster to build the seed.
- **Source entropy (*S*-H):** a source classifier is used to calculate the target instances' entropy score and select the ones with highest uncertainty as seed. Both language dependent (S_D) and independent (S_I) models are tested.
- ***S*-H and K-MC Combination:** the top ranked instances are selected as seed according to their entropy and K-MC combination score following Equation 1.
- ***S*-H and Length Combination:** the instances are ranked according to their entropy and length combination score and the top ones are chosen as seed. The combination score of an instance is the product multiplication of its rescaled (range 0-1) length with its entropy score.

Table 2. F1 and standard error results for different seed selection methods and sizes

Method	**B=100**	**B=250**	**B=500**	**B=1000**
Random	0.598 ±.024	0.735 ±.013	0.793 ±.004	0.829 ±.004
Length	0.749	0.811	0.834	0.851
K-MC	0.655 ±.016	0.8 ±.004	0.832 ±002	0.864 ±001
K-MC & Length	0.665 ±.021	0.794 ±.003	0.831 ±.004	0.864 ±.001
S_D-H Sum	0.746	0.809	0.838	0.845
S_D-H Mean	0.108	0.456	0.627	0.66
S_D-H K-Max	0.717	0.762	0.828	0.854
S_D-H Max	0.69	0.737	0.783	0.849
S_I-H Sum	0.777	0.806	0.831	0.862
S_I-H Mean	0.08	0.289	0.432	0.609
S_I-H K-Max	0.769	0.797	0.821	0.863
S_I-H Max	0.67	0.762	0.807	0.862
S_I-H Sum & K-MC	**0.79** ±.005	**0.839** ±.003	**0.858** ±.002	**0.876** ±.001
S_I-H K-Max & K-MC	0.756 ±.007	0.805 ±.005	0.845 ±.002	**0.879** ±.002
S_I-H Sum & Length	0.77	0.786	0.842	**0.878**
S_I-H K-Max & Length	0.766	0.786	0.842	0.873

Table 2 shows the results obtained for each explored configuration using practical seed sizes for a real environment. For the K-Means and Random selectors, their mean and standard error over 5 iterations are shown. As expected, the random baseline performs the worst. Selecting instances according to their length gives good results, although performance decreases as the number of selected instances increases. The K-MC sampling method yields better results as the batch size increases, demonstrating that input diversity plays an important role for instance sampling. Nevertheless, results for smaller seeds are lower than using other methods because information content is not taken into account. When transferring knowledge from the source classifier using the entropy scorers Sum and K-Max[2], S_I models perform slightly better than S_D models in smaller seeds, although such difference gets narrower in bigger seeds. While both S-H K-Max and S-H Sum demonstrate to be useful, the latter has direct relation with instance length, as the longer it is the more likely it is to have a higher entropy sum. It also shows similar patterns to the utterance length method, no longer being among the top methods in the largest seed size tested. The reason for this could be that it takes into account all the words of the instance, thus being sensitive to noise. On the other hand, the S-H K-Max scorer takes into account only the K words with highest entropy of the instance so it is more agnostic to length and low-entropy words in the utterance, making it more robust to noisy instances. As S_I models yield better results in general, only this method was combined with length and KMC.

The best results for seed selection are rendered by combining the K-MC method with the S_I-H Sum scorer, as this method takes into account the divergence between the input data, the length of the input samples, and their uncertainty. Likewise, the combination of K-MC with the S_I-H K-Max model yields better results as the seed size increases. It is interesting to note that when K-MC is combined with S-H the standard error intervals are reduced, improving the robustness of the method.

4.2 Query Selection Strategy Evaluation

In this section, the AL process is evaluated using different query strategies. To visualise the impact on learning speed, the learning curves for the base learners trained on the first 10.000 selected samples of the target corpus are plotted in Figure 2. The classifiers were asked to stop learning when they reached top performance[3]. The best configuration of Section 4.1 is used as seed. The query strategies evaluated are: Random baseline (R), Target domain T-H Sum/K-Max, the product multiplication of T-H Sum/K-Max with S_I-H Sum/K-Max, and the K-MC combination with T-H Sum/K-Max following Equation 1. The selected size for both seed and query batches was 250 instances.

[2] K=3 is used throughout the experiments

[3] Instead of using the hard 0.935 top performance score, a minimally softened breakpoint of 0.9345 was set.

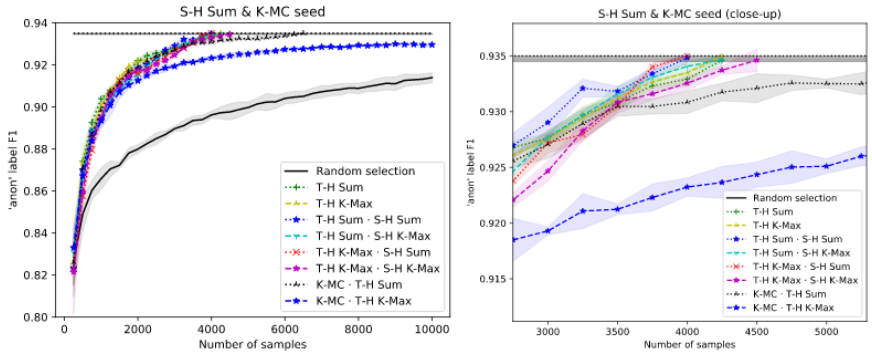

Fig. 2. F1 results of using the different query strategies for AL (left), and close up look of the top performance reaching iterations of the best methods (right). The standard deviation over 5 iterations appears shadowed

The reported learning curves show that all the proposed query selection methods perform significantly better than passive random selection, showing a much steeper curve and reaching top performance in fewer iterations. The two methods which consider target domain information only (T-H Sum and T-H K-Max) perform equally well, reaching top performance trained on 4250 instances only. Methods which combine target and source model information are in the top positions of best possible query strategies. Although the former has a slightly less steep curve in the first iterations, both the ***T*-H K-Max · *S*-H Sum** and the ***T*-H Sum · *S*-H Sum** combinations reach the top score with just 4000 training instances (less than 10% of the target training corpus), outperforming methods which do not use source model information. On the other hand, the combination of target and source H K-Max scores performs moderately worse than the mentioned methods, even the ones which do not consider S-H. Considering that H K-Max is agnostic to instance length we may conclude that this aspect may actually play a somewhat important role in best query selection. This hypothesis is supported by the fact that K-MC combination with T-H Sum has better results than its combination with T-H K-Max as the number of instances in the training set gets larger, although the two combinations perform the worst among the AL strategies tested.

5 Conclusions and Future Work

In this paper, new methods combining Knowledge Transfer and Active Learning to approach the lack of available annotated data for textual anonymisation have been proposed and compared. This has been done taking advantage of existing resources from a different language and domain. Exploiting classifiers trained on the source data, we demonstrate that the learning process on the target data for

the anonymisation task at hand can be notably speeded up from the very first batch, or seed, given to the target classifier.

Different scoring methods considering input divergence, length, uncertainty, and their combinations have been tested for seed selection and as AL query strategy criteria. Best seeds were achieved using scorers that considered all three aspects. For query strategy, methods that combined information from the source and the target models were the ones which performed better. With such query strategy methods and best seed selection, top classifier performance was reached using less than 10% of the full training data.

As future work, we plan to test this methodology for non-binary classification tasks, and to explore new ways to exploit information from multiple source model classifiers from different domains for textual anonymisation and other tasks.

References

1. Álvarez, A., Martínez-Hinarejos, C.D., Arzelus, H., Balenciaga, M., del Pozo, A.: Improving the automatic segmentation of subtitles through conditional random field. Speech Communication 88, 83–95 (2017)
2. Baker, L.D., McCallum, A.K.: Distributional clustering of words for text classification. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 96–103. ACM (1998)
3. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine learning 15(2), 201–221 (1994)
4. Dias, F.M.C.: Multilingual Automated Text Anonymization. Master's thesis, Instituto Superior Técnico de Lisboa (2016)
5. Dligach, D., Palmer, M.: Good seed makes a good crop: accelerating active learning using language modeling. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 6–10. Association for Computational Linguistics (2011)
6. Dorr, D.A., Phillips, W., Phansalkar, S., Sims, S.A., Hurdle, J.F.: Assessing the difficulty and time cost of de-identification in clinical narratives. Methods of information in medicine 45(03), 246–252 (2006)
7. García-Sardiña, L., Serras, M., Pozo, A.D.: ES-Port: a Spontaneous Spoken Human-Human Technical Support Corpus for Dialogue Research in Spanish. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan (May 7-12 2018)
8. Kleinberg, B., Mozes, M., van der Toolen, Y., et al.: Netanos-named entity-based text anonymization for open science (2017), preprint on Open Science Framework
9. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
10. Lewis, D., Gale, W.: Training text classifiers by uncertainty sampling. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Springer-Verlag New York, Inc. (1994)
11. Li, S., Xue, Y., Wang, Z., Zhou, G.: Active learning for cross-domain sentiment classification. In: IJCAI. pp. 2127–2133 (2013)
12. Li, X.B., Qin, J.: Anonymizing and sharing medical text records. Information Systems Research 28(2), 332–352 (2017)

13. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No.14. pp. 281–297. Oakland, CA, USA (1967)
14. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014)
15. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. pp. 188–191. Association for Computational Linguistics (2003)
16. Medlock, B.: An introduction to NLP-based textual anonymisation. In: Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), Genes, Italy (2006)
17. Olsson, F.: Bootstrapping named entity annotation by means of active machine learning: a method for creating corpora. Ph.D. thesis, U. of Gothenburg (2008)
18. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22(10), 1345–1359 (2010)
19. Rai, P., Saha, A., Daumé III, H., Venkatasubramanian, S.: Domain adaptation meets active learning. In: Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing. pp. 27–32. Association for Computational Linguistics (2010)
20. Settles, B.: Curious machines: Active learning with structured instances. Ph.D. thesis, University of Wisconsin–Madison (2008)
21. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648 (2010)
22. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 6(1), 1–114 (2012)
23. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics (2008)
24. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the fifth annual workshop on Computational learning theory. ACM (1992)
25. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. pp. 134–141. Association for Computational Linguistics (2003)
26. Shannon, C.E.: A note on the concept of entropy. Bell System Tech. J 27(3), 379–423 (1948)
27. Shi, X., Fan, W., Ren, J.: Actively transfer domain knowledge. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 342–357. Springer (2008)
28. Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art anonymization of medical records using an iterative machine learning framework. Journal of the American Medical Informatics Association 14(5), 574–580 (2007)
29. Tomanek, K., Laws, F., Hahn, U., Schütze, H.: On proper unit selection in active learning: co-selection effects for named entity recognition. In: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing. pp. 9–17. Association for Computational Linguistics (2009)
30. Tomanek, K., Wermter, J., Hahn, U.: Efficient annotation with the jena annotation environment (JANE). In: Proceedings of the Linguistic Annotation Workshop. pp. 9–16. Association for Computational Linguistics (2007)