EXTRACTING INFORMATION FROM PARTICIPIAL STRUCTURES

Enikö Héja, Kata Gábor

Linguistic Institute, Hungarian Academy of Science and Eötvös Loránd University

Abstract

Our applied linguistic research aims at increasing the efficiency of a rule-based information extraction (IE) system by enhancing it with further grammatical knowledge. The input of the IE system is made up of sentences of business news. The event of the piece of news is identified through the main verb of the sentence, while participants and circumstances of the event through arguments and adjuncts of the main verb. Our objective was to unfold the hidden information, contained by NPs within which non-finite verbs (e.g. participles) appear. Thus, we invented a rule-system to transform participial structures into sentences with a finite verb, so that they could serve as input of the IE system. To tackle this task we had to be able to distinguish between real participles and adjectives. According to us there are some distributional criteria which can be used as the basis for creating the right classification.

1. Introduction

In what follows we would like to present our applied linguistics research which aims at increasing the efficiency of an already existing rule-based information extraction system by enhancing it with further grammatical knowledge. Our work concentrates on the NewsPro information extraction system (Prószéky 2003), developed jointly by MorphoLogic Ltd., Institute of Informatics at Szeged University, and Linguistics Institute of Hungarian Academy of Sciences. The system was developed and tested on a corpus of short business news.

Firstly, NewsPro performs a shallow syntactic analysis on the input text, then it matches pre-defined semantic patterns —so-called 'event frames'— to the text. In case of successful pattern matching, slots of event frames are filled by the elements of the text, thus the output identifies the main event of the piece of news as well as its participants and circumstances. Semantic patterns are centered around finite verbs while their complements and adjuncts represent participants and circumstances, respectively. Thus, pattern matching is based on the *finite* verb previously recognized as predicate, and its argument structure. This method relies on the supposition that in short news it is always the verbal predicate that expresses the main event. Although this approach proves to be working in most cases, it has the disadvantage of omitting secondary information (frequently indicated as the cause or the antecedent of the main event) from pattern matching. The reason is that secondary information is represented grammatically by non-finite verbal forms such as participles or deverbal nouns.

For example:

> '*[$_{part}$A cég által kedden meghozott döntés] nyomán sokan keresnek új munkahelyet.*'
> 'Due to [the decision made by the firm on Tuesday], many people are looking for a new job.'

In the sentence above, NewsPro is able to identify the main event (i.e. looking for new jobs), but not the bracketed constituent, which expresses an earlier event, conceived as the cause of the main event. However, the user may be interested to learn about the antecedents and the connection between the two pieces of information.

This phenomenon is supposed to be handled by a preprocessing module within NewsPro. Preparation of transformational rules as well as other tasks related to the preprocessing of the text were performed by Intex (Silberztein 1993). The module transforms input participial structures into complete sentences with a finite verb as their predicate. Further steps of the processing, such as syntactic parsing and semantic pattern matching may run on the transformed sentences without any modification.[1] Moreover, as Hungarian constituent order is relatively free, we expect the system to yield better results on automatically generated sentences, as their constituent order is homogeneously SVO.

2. The Corpus Annotation Tool

The implementation and testing of the transformation rules, as well as any task the preprocessing of the text involved were carried out using Intex, a powerful corpus processing tool freely available for research purposes. Intex is particularly suitable for implementing lexicalist approaches to language processing, as it makes wide use of several types of structured dictionaries. The feature we took particular advantage of is that morphosyntactic and semantic description of words are available at every level of the analysis. This allowed us to create transformation rules which referred to the base verb of participles (that we also coded in the dictionary), and the semantic-syntactic properties of the base verb.

3. The Outlines of the Problem

The success of the preprocessing module on one hand depends on the *grammatical and semantic well-formedness* of the output (theoretical requirement) and on the other hand on *the degree of informativity* of the transformed sentences (practical requirement). We made an attempt to elaborate an algorithm for filtering out supposedly informative participial structures on the basis of solely grammatical information. Below we give a brief description of the method we used.

[1] We have not dealt with the morphological aspect of the elements in the resulting sentences yet, the morphological module of Hungarian Intex is presently under development.

3.1. Participles and Participial Structures in Hungarian

The focus of our research consists in past participles. Hence, we need to introduce the main features of Hungarian past participles. Although linguists did not reach a consensus about the exact status of participles in Hungarian, it is widely accepted that participles originate from verbs, either by derivation or by inflection. From our point of view the precise nature of this process plays no role, in tandem with the consideration that there might be no strict boundary between derivational and inflectional suffixes. Past participles can be derived freely from verbs if the verb has an argument the thematic role of which is patient or theme.

Below we present the form of Hungarian past participles:

Verb - (Vowel) - (t) - t

The form of Hungarian past participles coincides with the past tense form of the corresponding verb. Some examples: *ad - ott* ‘given’; *megérdemel - t* ‘deserved’.

The expression of the form *verb - suffix* is the head of the participial structure. The characteristics of participial structures which enable the production of well-formed sentences by means of our rules are on one hand that the participle preserves the meaning of its base verb, and on the other hand that the arguments of the base verb can be derived from the internal structure of the NP containing the given participial structure. As the internal structure of a Hungarian NP is rather strict, our rules are able to recognize the constituents of it and identify them as adjuncts and complements of the base verb.

3.2. The Problem

First, let us have a look at some examples, which might be the output of our transformation rules, but are not able to serve as input of the pattern matching process:

‘a *jegyzett* tőke’	[particip Valaki jegyzett tőke -t]
The subscribed capital	Somebody subscribed capital - *ACC*
‘a *nyomott* hangulat’	[particip Valaki nyomott hangulatot -t]
The depressed mood	Somebody depressed mood - *ACC*
‘a *mérsékelt* PC-chip kereslet’	[particip Valaki mérsékelt PC-chip kereslet-t].
The moderated PC-chip demand	Somebody moderated PC-chip demand - *ACC*
a *nyomtatott* sajtóban	[particip Valaki nyomtatott sajtóban -t]
The printed media - *INE*	Somebody printed media - *ACC*
‘a *ragozott* szóalakok - ból’	[particip Valaki ragozott szóalakokból -t]
The inflected word forms - *ELA*	Somebody inflected word forms
‘a *kerekített* euróár - ak’	[particip Valaki kerekített euróárak -t]
The rounded Euro price - *PL*	Somebody rounded Euro prices - *ACC*
‘a *használt* ingatlan - ok’	[particip Valaki használt ingatlanok -t]
The used property - *PL*	Somebody used properties - *ACC*

The sentences above are in some way semantically ill-formed. For instance in the second example *nyomott hangulat* ‘depressed mood’ has nothing to do with the fact that somebody depresses something. There are also cases which are not as bad as this.

Taking *ragozott szóalakok* 'inflected word forms', we find that the resulting sentence is improper in another way. Namely, this NP denotes rather a state than the process itself, so *ragozott szóalakok* does not mean that there is somebody who really did inflect those word forms, instead it means that those word forms are in a certain state, i. e. they have certain suffixes.

So, we have to be able to tell apart participial structures that result in well-formed sentences from those that do not. Our solution is based on the fact that there is an adjective-participle homonymy in Hungarian. Actually, we state that those transformations that contain *adjectives* produce ill-formed sentences, and the structures containing *participles* can serve as the input of our rules.

3.3. Adjectives and Participles

As we mentioned above there is no consensus among Hungarian linguists about the exact nature of participles. One fact which is responsible for this insecurity is homonymy between adjectives and participles. As some authors state, there are certain cases when the nature of a given expression is absolutely undecidable (e.g. *kedvelt*, 'much liked') (Komlósy 1992 and Kenesei 2000). Thus, the question arises: on what ground is it possible to distinguish between adjectives and participles? We follow Komlósy's (Komlósy 1992) suggestion according to which there are syntactic —mainly distributional— tests, which we can rely on to make our decision.

Before we list them, we have to note that semantic facts also support our hypothesis above. Namely, as participles keep the main characteristics of a verb —for example its argument structure and the ability to place the time of the event denoted by the participle in relation to the time of the main verb of the sentence (Kiefer 2000)— they are capable of preserving the event structure of the original verb, as opposed to adjectives. According to us this is the most important feature of participles which guarantees the well-formedness of the resulting sentence.

Now, coming back to our main train of thoughts we list the syntactic tests, mentioned above:

(1) *comparison*: only adjectives can undergo comparison,
(2) *deriving adverbs*: only adjectives can serve as input of adverbial derivation,
(3) *predicative use:* only adjectives have predicative use,
(4) *preverb detachment:* only preverbs in participles could be detached when there is a negation.

Unfortunately, the criteria above are not able to help us directly in telling apart adjectives from participles, since we cannot take the possible transformation of the texts' elements into consideration when applying our rules.

3.4. Our Solution

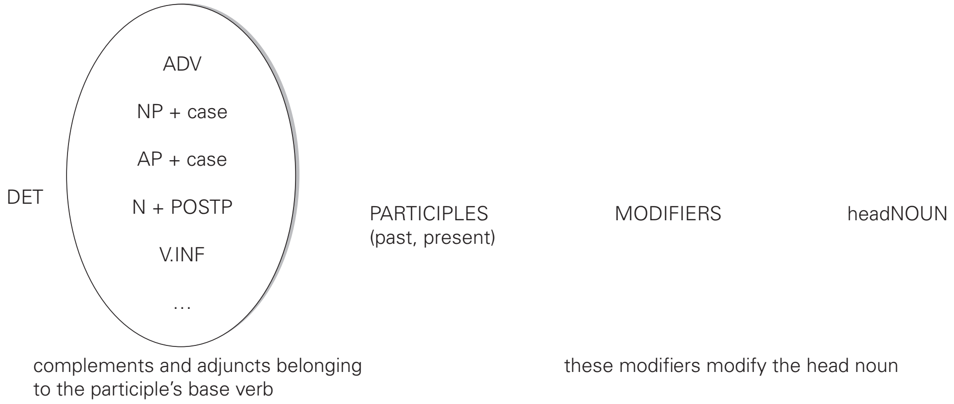
Thus, the question remains: how could we distinguish between adjectives (the transformations of which supposedly result in ill-formed sentences) and participles (with which the rules output well-formed sentences) *in situ*?

We found that the three criteria below are sufficient to make the right classification:

(1) If at least one of the base verb’s *complements* is present, then it is a participle (Komlósy 1992).
(2) If at least one of the base verb’s *adjuncts* is present, then it is a participle.
(3) If at least a *preverb* is present, then it is a participle.

The acceptability of the (1) criterion follows from the fact that participles are “closer” to the verb from a derivational aspect and they keep the arguments of the base verb (or at least some of them.)

To support the (2) criterion, we have to elaborate the statement above according to which the internal structure of the Hungarian NP is quite bound. The following illustration represents the constituent order within the NP.


ADV
NP + case
AP + case
N + POSTP
V.INF
…
DET
PARTICIPLES
(past, present)
MODIFIERS
headNOUN
complements and adjuncts belonging to the participle’s base verb
these modifiers modify the head noun

This means that the constituents within the ellipse are always attached to the participle, because they were complements and adjuncts of the base verb, as the following example illustrates:

‘*[$_{part}$A cég eladás - a kapcsán felmerül - t] hamis [$_{N}$híresztelés]*’
DET firm sale - *POSS* in-connection-with emerge - *PAST* false rumor
‘A false rumor that emerged in connection with the selling of the firm’

In this NP *‘A cég eladása kapcsán’* is an adjunct of the base verb, while *‘hamis’* is a modifier of the head noun *‘híresztelés’*.

The (3) condition correlates with the observation that while participial phrases put emphasis on the course of events denoted by the base verb, adjectives express states of processes. As Hungarian preverbs’ main function is to express aspect, it is conceivable that the presence of a preverb supports the ‘course of events’ reading, which means that there is a participle. Beside these intuitions, the above mentioned distributional facts also supply confirmation for our hypothesis.

(1) comparison:
**’a [tegnap mérsékel]-t-ebb kereslet’*
the [yesterday moderate]- *SUFF- COMP* demand

(2) predicative:
*'*Ez a szóalak [az órán ragoz]-ott.*'
this word form [during the class inflect]- *SUFF*

(3) ADV formation:
'[1000 Ft-ra mérsékel]-t-en*'
[1000 Ft-onto reduce]- *SUFF- ADV*

(4) preverb detachment:
*'*A házak [fel nem újít]-ott-ak.*'
the houses [re- not store]- *SUFF - PL*

Consequently, our main hypothesis is proved: only those expressions which do have complements, adjuncts or preverbal modifiers are participles, and those which have neither of them in their left context are adjectives. Since our original aim was to filter out structures bearing information, we have to examine how syntactically separated groups of participles and adjectives can be compared with the informative-uninformative partition. In fact, it turns out that informative structures coincide with participial structures.

4. The Grammar

When setting up our transformational rules, we used the following presuppositions:

(1) it is possible to derive participles both from transitive and intransitive verbs,
(2) if the verb is intransitive the head of the noun phrase is the subject of the participle's base verb,
(3) if the verb is transitive the head of the noun phrase is the object of the participle's base verb,
(4) the complements and adjuncts of the base verb appear before the participle,
(5) as the past participle usually expresses anteriority in time, the transformed finite verb is in past tense.

In addition, we dealt only with NPs which begin with a determiner. We decided to do so because complements and adjuncts appearing before participles might be extremely diverse, which makes the exact recognition of the NP's left boundary hard. Using a determiner as the left boundary of the NP enables us to identify all constituents between the participle and the beginning of the NP as complements and adjuncts of the base verb, while expressions appearing between the participle and the head noun can be identified as the modifiers of the noun (in accordance with the illustration above).

Firstly, on the basis of the (1)-(3) conditions we divided the NPs into two groups: one of them consists of participles derived from transitive verbs and the other from intransitive verbs. The practical purpose of that was to encode the feature of transitivity in a dictionary, by means of which local grammars are able to perform the transformation.[2]

[2] We used the verbal argument structure database prepared by Corpus Linguistic Department to develop our dictionary (i. e. to encode the relevant syntactic characteristics of verbs).

4.1. Transitive Verbs

We considered a verb transitive if it had at least one transitive occurrence in our database. To transform this kind of participial structures into finite sentences we used the following algorithm:

Det (V_compl/V_adj) V_{past_part} N → Valaki V_{past_part} Det N - ACC (V_compl/V_adj)

The rule above says that if an NP consists of a sequence of a *determiner* 'Det'), *complements* and/or *adjuncts* of the base verb ('V_compl/V_adj'), a *past participle* ('V_{past_part}') and a *head noun* ('N'), it has to be transformed into a string which is made up of *Valaki* (the Hungarian counterpart of 'somebody'), the *past participle*, a determiner, the head noun with accusative case and finally, the complements and/or adjuncts of the base verb. Here parentheses mark optionality. The reason for the acceptability of the past participle's use at the right hand side of the rule is the fact that in Hungarian the forms of past participles and the forms of the corresponding past tense verbs coincide. The transformation is illustrated by the example below:

'*[$_{part}$A bővítés után - ra tervez - ett] munkaerővándorlási [$_{N}$korlátozás]*'
DET expansion after - *SUB* plan - *SUFF* migration-of-labor restriction
'The restriction of migration of labor planned to be introduced after the expansion'

'*Valaki tervezett munkaerővándorlási korlátozást*
'Somebody planned to introduce the restriction of migration of labor
bővítés utánra.'
after the expansion.'

In other words, the first step of filling the slots in the argument structure of the base verb is to identify the head of the NP as the object of the resulting sentence. Secondly, the subject position —since in most cases the subject itself does not appear in this kind of participial structures— is occupied by the expression *Valaki* 'Somebody'. This solution is made possible by the fact that in these cases the subject is usually an agent. Actually, though not frequently, the subject might also appear in the participial structure. When this happens, it could be expressed by the postposition *által* 'by'. Such structures are handled by the following rule:

Det N_{subj} által (V_compl/V_adj) V_{past_part} N → N_{subj} V_{past_part} N - ACC (V_compl/V_adj)

As in the example below:

'*[$_{part}$A budapest - i cég által rendszeresen közzéte(sz) - tt] eredmény - ek*'
DET Budapest - *ADJ* firm by regularly publish - *SUFF* results - *PL*
'Results published by the firm in Budapest regularly'
'*A budapesti cég közzétett eredményeket rendszeresen.*'
'The firm in Budapest published results regularly.'

There are cases, when the subject of the base verb appears in the participial structure, but it is not expressed by an *által* postpositional phrase. In such cases the NP denoting the subject of the base verb is in nominative case morphologically, and it is the possessor of the head noun of the main NP. However, this construction does not contradict our hypothesis, according to which the constituents preceding the participle are complements and adjuncts of the base verb, since the possessor plays the role of the subject in the transformed sentence:

‘[$_{part}$*A svéd Networks tervez - ett] adósságátalakítás - i program - já - ban*’
DET Swedish Networks plan - *SUFF* debt-conversion - *ADJ* program - *POS* - *INE*
‘In the debt conversion program planned by the Swedish Networks’
‘*A svéd Networks tervezett az adósságátalakítási programot.*’[3]
‘The Swedish Networks planned the debt-conversion program’

4.2. Intransitive Verbs

We considered a verb as intransitive if our lexical database did not contain any transitive occurrence of it. Regaining the original argument structure, i.e. that of the base verb, was quite simple: the head of the NP is identified as the transformed sentence’s subject. Just as in the case of transitive verbs, complements and adjuncts of the base verb precede the participle. In the resulting sentence they follow the finite past tense verb. It might be interesting that the subject of base verbs belonging to this class is usually a patient.[4] We used the rule below to transform such structures:

Det (V_compl/V_adj) V$_{past\_part}$ N → Det N V$_{past\_part}$ (V_compl/V_adj)

‘[$_{part}$*A bécs - i kereskedelmi bíróság - on tegnap lezajl - ott] tárgyalás*’
*DET*Vienna - *ADJ* mercantile court - *SUP* yesterday pass-off - *SUFF* trial
‘The trial which passed off yesterday at the mercantile court in Vienna’
‘*A tárgyalás lezajlott tegnap a kereskedelmi bíróságon Bécsben.*’
‘The trial passed off yesterday at the mercantile court in Vienna.’

As the example shows, the argument structure of intransitive verbs can be fully reconstructed from the NPs internal structure. Nevertheless, we have to note that intransitive verbs are less useful regarding information extraction. This is because they express less implicit information, since these participles are usually derived from verbs with only a vague semantic content. For instance: *bekövetkezik* ‘come true’, *beindul* ‘start up’, ‘be launched’, *létrejön* ‘come into existence’, *kialakul* ‘take shape’. Hence, the identification of their arguments adds probably no extra information to our existing knowledge. Still, such structures could be worth dealing with, since they might help us in bringing hidden relations to light.

5. Evaluation

We scrutinized a total of 7058 sentences, i.e., 43% of the whole corpus. As the corpus is split up in smaller texts according to their topic, we decided to evaluate

[3] Unfortunately, some occurrences contradict this consideration. Occasionally we can only rely on our knowledge of the world when deciding whether the head noun’s possessor is equal to the subject.

[4] There is an other use of this structure, the so-called newspaper language use. In such cases the subject of the base verb may be also an agent. (e.g. ‘[part A tegnap lemondott] elnök’-‘The president, who resigned yesterday’). This use lays emphasize on the anteriority in time (Laczkó 2000).

the results on a set of randomly selected 500-sentence long fragments to avoid the useless repetition of the same patterns and the eventual lack of different kinds of structure.

The number of hits in the test corpus is 798, with a precision of 64%. We were not able to count recall values as there is no manually annotated Hungarian corpus which contains annotation of NP-internal participial structures. However, as this application is supposed to improve the efficiency of an IE system, where correctness is more important for the users than the amount of the output, we have reasons for focusing on precision.

Most of the improper hits are due to one of the following deficiencies:

1. Improper morphological analysis, hence inaccuracies in the dictionary are responsible for most of the false hits.
2. As it has been already mentioned, we cannot handle NPs without a determiner. However the more informative (and at the same time longer) NPs relatively often begin with a determiner.
3. Ensuing from the nature of the texts there are a lot of peculiar nouns in the corpus. Unlike a usual Hungarian text, the occurrence of firms' and trades' names is quite common, such us compound words with NN internal structure, the recognition of which is also difficult.
4. There are lexicalized participles, too. Although they might have preverbs, adjuncts or complements in their context they behave as adjectives (e.g. *elmúlt* 'bygone', *ismert* 'known').

As it is obvious from the error types, this rule system is a high-level language processing application, where most of the errors are due to problems of the lower level analysis such as morphology, dictionaries or named entity recognition. On the other hand, as it is a rule-based application, the deficiencies that its evaluation brings into light can be directly turned into useful information for outlining the directions of future development.

References

Kenesei, I., 2000, «Szavak, szófajok, toldalékok». In: *Strukturális magyar nyelvtan 3.* ed. by Kiefer, F. pp. 75-135. Budapest. Akadémiai Kiadó. ('Words, Lexical Categories, Suffixes'. In: Structural Grammar of Hungarian 3.)

Kiefer, F., 2000, *Jelentéselmélet*. Budapest. Corvina ('Semantics').

Komlósy, A., 1992, «Régensek és vonzatok». In: *Strukturális magyar nyelvtan* 1. ed. by Kiefer, F. pp. 299-529. Budapest. Akadémiai Kiadó ('Predicates and Complements'. In: *Structural Grammar of Hungarian* 1.)

Laczkó, T., 2000, «A melléknévi és határozói igenévképzők». In: *Strukturális magyar nyelvtan 3.* ed. by Kiefer, F. Budapest. Akadémiai Kiadó. ('The Suffixes of Adjectival and Adverbial Participles'. In: *Structural Grammar of Hungarian* 3.)

Prószéky, G., 2003, «Information Extraction from Short Business News Items». In: *Proceedings of the First Hungarian Computational Linguistics Conference*, ed. by Alexin, Z. and Csendes, D. pp. 161-167. Szeged. Szeged U. P.

Silberztein, M., 1993, *Dictionnaires électroniques et analyse automatique de textes: Le systeme Intex*. Paris. Masson.

Vajda, P. et al., 2004, «A ragozási szótártól a NooJ morfológiai moduljáig». In: *Proceedings of the Second Hungarian Computational Linguistics Conference*, ed. by Alexin, Z. and Csendes, D. pp. 183-190. Szeged. Szeged U. P.