Creating Syntactically Annotated Romanian Corpus

Victoria Bobicev
Faculty of Computers, Informatics and Microelectronics
Technical University of Moldova
vika@rol.md

Keywords: corpus, morphological annotation, chunking, grammar rules creation, prepositional phrase attachment

Abstract

This paper presents an attempt to create a syntactically annotated Romanian corpus. Several steps of annotation are described. Firstly, we solve the problem of dividing a text into sentences using machine-learning method. Then each word is morphologically annotated. For morphological disambiguation we apply rule-based method. The third step is shallow parsing. Unification-based grammar rules were semi-automatically created for chunking. Groups of words and their characteristics were extracted and sorted automatically by frequency of occurrence and the obtained list that represented frames for grammar rules was checked manually. We applied those rules for automatic chunking. The similar method was applied for extracting Named Entities. Named Entities are annotated in texts before chunking. The last step described in paper is an attempt to solve the problem of chunks attachment.

The texts are annotated in XML according to standardization proposed in MULTEXT project.

1 Introduction

Nowadays a linguistically annotated corpus is a necessary instrument in the area of natural language researches. Corpus is a widely used tool for teaching materials, classroom exercises and research purposes in the field of computational linguistics. A large number of corpus creating projects have recently been carried out [Bulric-Ling]. However, today there is no a syntactically annotated Romanian corpus for the statistical methods use. The Romanian corpus we know created as a result of MULTEXT-EAST project contains only the morphological annotation and is not freely accessible [MULTEXT-East Specifications].

Our aim is to create the syntactically annotated corpus of Romanian texts.

Description of the corpus texts. Our task was to find a great volume of texts with diacritics written in Moldova. In our case we obtained law documents posted on the site http://moldova.wjin.net as a result of "World Justice Information Network Moldova (WJIN Moldova)" project. Though the texts placed on the site have no diacritics, we have received texts before the diacritics were removed. SGML encoding method is used to keep diacritics [MULTEXT-East Specifications]. These texts represent obviously expressed sublanguage that has partly facilitated their analysis.

The paper describes the following annotation stages of our corpus [Bulric-Ling]:

- initial segmentation of document's structure;
- morphological annotation;
- chunking;
- preposition phrase attachment.

2 Splitting the text into sentences

It is known that even such a trivial task as splitting a text into sentences creates a number of difficulties for an automatic system. They are most obvious in the example of our texts:

Părţile la 7.02.97 şi 10.03.97 au încheiat contractele nr.nr. 21 şi 28 în suma de 7.345 lei.

(Parties signed the contracts nr. Nr. 21 and 28 for 7.345 lei on 7.02.97 and 10.03.97)

As it is seen, there are 7 periods and only one full stop in the example.

Applied method. In our case it was decided to use machine-learning method [Ratnaparkhi, 1998]. For this purpose, in the texts containing about 100.000 words sentence boundaries were manually marked. A set of features containing information about the period and its neighborhood was selected automatically from these texts. Words in front of the period and after it and certain word peculiarities (numbers or capital letters) were taken. For each feature the probability was calculated.

The program based on this list calculates the probability for each period to define the end of the sentence. The decision was made on comparing two calculated probabilities and choosing the variant with the greatest probability.

The system work results are low enough, about 90%, but for our texts even this percentage improves the text annotation, as there is a considerable amount of abbreviations and figures, containing periods that are not full stops. Without this program about 25-45% of periods were wrongly determined as the sentence boundary. The result of program can be improved if a greater volume of texts is taken for training.

3 Morphological annotation

The next step is the lexicon-based morphological annotation. The text is split into words and other lexical units (punctuation marks, figures) and each word is given its equivalent morphological features from the dictionary.

Dictionary. The morphological dictionary we had at our disposal was created several years ago by scanning the DEX [DEX, 1996]. Each line of the formed dictionary contained one lemma, its initial form and its morphological description. We have changed morphological encoding according to EAGLES and MULTEXT recommendations [MULTEXT Specifications]. A morphological tag contains letters, each of them being one morphological feature of the word.

In the article about the creation of Romanian lexicon during the MULTEXT-EAST project, the total number of 614 MSD (morpho-syntactic descriptions) was presented [Tufis, 2000]. We reduced this number to 138 as we had used a smaller number of characteristics for each part of speech [Tomaz Erjavec et al., 1998].

Methods used for disambiguation. Having some volume of the corpus checked manually, we used Marcov model for morphological disambiguation. In applying it for our texts, however, we were faced with the same problem as was described in [Tufis, 2000]. Even for 138 tags the matrix obtained for trigram-based HMM is too large. Therefore, we have taken only 12 basic characteristics of parts of speech. As the majority of unknown words were proper names and abbreviations, Viterbi algorithm was supplemented by a few rules determining them. The result of the algorithm application was good, however, only for the basic characteristic of the word: V - verb, N – noun, Aj – adjective and others. This information is not enough for further syntactical text annotating.

Therefore, to define characteristics of unknown words, we have created the array including word inflexions and morphological tags corresponding to them. Having found the unknown word, the program addresses the inflexion array and puts forward all the matching morphological codes. Then empirical transformational-based rules were applied to text for the morphological disambiguation. The sample of the rule is:

if *the word after word "a" is an infinitive* **then** *"a" is a particle*

Both the statistical method on the basis of HMM and rule-based method have resulted in about 95-97% correctly marked words. Thus, the texts have to be checked manually.

In accordance with standardization recommendation [MULTEXT Specifications], XML has been used for annotation. In Fig. 1 a fragment of annotated text is presented.

4 Chunking

Since the full syntactic analysis is a very complicated task, we decided to implement chunking as the first step [Abney, 1996b]. This splits sentences into noun phrases, verb phrases, prepositional phrases etc. We considered that the simplest way of chunking is the use of regular expressions based on morphological information. To find noun phrases, word groups containing nouns, adjectives, pronouns, numerals, articles, determinants, prepositions and conjunction were searched. Verb phrases were considered as groups of verbs, adverbs and particles. A distinctive feature of the Romanian language is the fact that verb phrase often has a reflexive pronoun, which relates to it.

```
<sentence>
      <word id=”745” infin="plenul" part="Ncmsry">Plenul</word>
      <word id=”746” infin="curte" part="Ncfsoy">Curţii</word>
      <word id=”747” infin="suprem" part="Afson">Supreme</word>
      <word id=”748” infin="de" part="S">de</word>
      <word id=”749” infin="justiţie" part="Ncfsrn">Justiţie</word>
      <sign>,</sign>
      <word id=”750” infin="casa" part="Vmg">casând</word>
      <word id=”751” infin="hotărâre" part="Ncfpry">hotărârile</word>
      <word id=”752” infin="instanţă" part="Ncfpoy instanţelor</word>
      <word id=”753” infin="judecătoresc" part="A-poy">judecătoreşti</word>
      <word id=”754” infin="cu" part="S">cu</word>
      <word id=”755” infin="remitere" part="Ncfsry">remiterea</word>
      <word id=”756” infin="cauză" part="Ncfsoy">cauzei</word>
      <word id=”757” infin="pentru" part="S">pentru</word>
      <word id=”758” infin="rejudecare" part="Ncfsrn">rejudecare</word>
      <sign>,</sign>
      <word id=”759” infin="avea" part="Va--3s">a</word>
```

Fig.1. Fragment of morphologically annotated text

Prepositional is considered to be the noun phrase beginning with preposition.

When finding noun phrases from these texts two kinds of mistakes were noticed. If a comma separates the similar elements in one phrase, they are considered to be different phrases (ex. 1) and vice versa, two noun phrases that are not separated by any marks or verbs are considered to be one phrase (ex 2.).

privatizarea clădirilor, construcţiilor, încăperilor (ex. 1)
(buildings, constructions, premises privatization)

care motive reclamanta (ex. 2)
(what reasons plaintiff)

Extracting noun phrases including prepositions, 76% of noun phrases were found correctly. After excluding prepositions from noun phrases, the percentage of wrongly unified phrases decreased considerably. The success rate of this second variant was about 82% correctly extracted noun phrases. Because the texts were taken from a limited domain, verb phrases are even less various then noun phrases and don’t create difficulties on determining.

5 Grammar rules creation

We decided to create grammar rules to obtain better results of chunking.

Grammar rules were created in semi-automatic way [Bobicev, 2003]. The program extracted all the word groups from the texts as it was described above. Then all the extracted groups were sorted out by the frequency of occurrence and repeated groups were combined. Then only structures were taken and again the repeated ones were combined. Occurrence frequencies for structures have been calculated and sorted out. Deduced in this way, the list of structures was the base for creating grammar rules.

A simple context – free grammar is not the best way for formalizing such a language as Romanian because of the relatively free word order. Another peculiarity of Romanian is that it is a highly inflected language.

Therefore, we consider the grammar based on the unification [Shieber, 1986] to be the most suitable for formalization of Romanian. The main characteristic of the given type of grammar is the creation of two parallel structures. The former is a classical one and creates a tree-type phrase structure and the latter participates in forming a structure with the account of word characteristics and special categories in their linking.
The basic operation of these grammars is unification, i.e. the operation of joining structures into a common one that contains information from both unified structures [Covington, 1994].

There were words with morphological characteristics only at our disposal; they were used for determining constrains in the process of making up the rules.

The results of the program work that defines noun and verb phrases on the basis of formulated rules in a morphologically annotated text are good enough and there are about 92 % of correctly defined groups.

A fragment of annotated text after this stage is presented in Fig. 2.

Using similar technique, we have also created the program which annotates the Named Entities as there is a lot of mentioned persons, organizations, localities, dates, numbers of articles, etc. Named Entities are annotated before the chunking.

```
< sentence id="8">
  < chunk id="8_1" type="NP" >
    <word id=”346” infin="reclamanta" part="Ncfsry">Reclamanta</word>
  </chunk>
  < chunk id="8_2" type="VP" >
    <word id=”347” infin="se" part="Px3">s-</word>
    <word id=”348” infin="avea" part="Va-3s">a</word>
    <word id=”349” infin="adresa" part="Vmp-sm">adresat</word>
  </chunk>
  < chunk id="8_3" type="PP" >
    <word id=”350” infin="în" part="S">în</word>
    < chunk id="8_4" type="NP" >
      <word id=”351” infin="judecată" part="Ncfsrn">judecată</word>
    </chunk>
  </chunk>
```

Fig. 2 Fragment of chunked text.

6 Prepositional phrase attachment

As about 2/3 of noun phrases in text are dominated by prepositions, the next step can be considered as a problem of prepositional phrase attachment. Prepositional phrase attachment is a subtask of a general natural language problem. It is the task of choosing the attachment word of a preposition that corresponds to interpretation of the sentence. Though the necessity of world knowledge for correct prepositional phrase attachment is proved, this problem is more or less successfully solved using statistical or corpus-based approaches. Most of the corpus-based approaches consider prepositions whose attachment is ambiguous between a preceding noun phrase and verb phrase [Steina and Nagao, 1997]. It is a kind of classification task in which the goal is to predict the correct attachment given the head noun, the head verb, the preposition, and optionally, the object of the preposition. In English the attachment word of a preposition is usually located only a few words to the left of the preposition and it is either the nearest verb or the nearest noun [Ratnaparkhi, 1998].

The preliminary estimation of the consulted linguists was that approximately 78% of prepositions were attached to the nearest word.

We annotated manually prepositional attachment in chunked text of about 10 000 words. As a result of the annotation we obtained the following statistics.

Total number of preposition: 1866

Number of prepositions that are attached to the nearest word: 52%.

Number of prepositions that are attached to:

- noun 47%
- verb 42%
- adjective 6%
- pronoun 3%

As another result of our investigation we observed that the distance between preposition and attachment word may be of 6 or even more words and often the choice has to be made among several nouns.

In order to solve the problem of prepositional phrase attachment we made three experiments based on statistical method described in [Ratnaparkhi, 1998]. The method includes several steps. First, the training data is generated from the text annotated morphologically. Second, the statistical model is applied. The last step is disambiguation of prepositional phrase attachment in the unseen text.

In the first experiment we tried to use all our texts (about ½ million words) for extracting training data without any morphological information. In the second one we used morphologically annotated texts (100 000 words) and some heuristic rules for extracting data. The third variant used texts with manually annotation of prepositional phrase attachment (10 000 words). As a result we obtained three lists of word triples. The first word in triple is the word the preposition is attached to, the second is preposition itself and the third is the object of preposition. Then statistical model was applied for each of these lists, and the same text of 1000 words was annotated using three separate statistics. Having 134 prepositions in the test set we obtained results presented in Tab. 1.

Tab. 1. Results of prepositional phrase attachment experiments.

Experiment	Number of correct annotated prepositions	Per cent
1	97	72%
2	102	76%
3	110	82%

As it is seen the best result is obtained on the base of manually annotated texts, though other methods allow as improving the attachment in comparison with the baseline 52%.

7 Conclusions

In this paper we presented an ongoing work that creates a syntactically annotated corpus of Romanian texts. For all consecutive steps of text processing separate programs were developed:

- a program that defines sentence boundaries;
- a program that executes morphological annotation of words;
- a program that annotates Named Entities;
- a program that carries out partial parsing-chunking;
- a program that attaches prepositional phrases.

In spite of the fact that the results of programs are not the best, their application, nonetheless, allows to receive a Romanian text with partial syntactical tagging that further may be checked and edited manually. Even having such a corpus processed only automatically and not corrected by hand, one can use it for training many statistical methods based on noisy data. We hope to improve annotation methods in the process of further corpus creation and solve the problem of full parsing. We also consider that created programs can be used in other fields for Romanian language processing.

The annotated texts will be made available on a Website. We also hope to present demo-version of our programs on the same site.

References

[Abney, 1996a] Steven Abney. 1996. Tagging and Partial Parsing. In: Ken Church, Steve Young, and Gerrit Bloothooft (eds.), *Corpus-Based Methods in Language and Speech.* Kluwer Academic Publishers, Dordrecht. 1996.

[Abney, 1996b] Steven Abney. 1996. Partial Parsing via Finite-State Cascades**.** In *Proceedings of the ESSLLI '96 Robust Parsing Workshop.* 1996.

[Bobicev, 2003] Victoria Bobicev. 2003. Metoda semiautomata de creare a gramaticii pentru analiza sintactica a grupurilor nominale in textele Romanesti. *In Proceedings of International Conference Trends in the Development of the Information and Communication Technologies in Education and Management*, pages 301-305, Chisinau, Republic of Moldova, March.

[Bulric-Ling] Corpora and HLT. Current trends in corpus processing and annotation. Available at http://www.larflast.bas.bg.balric

[Covington, 1994] Michael A. Covington. 1994. Natural Language Processing for Prolog Programmers. Chapter 5. Unification-Based Grammar, pages 102-150. Prentice Hall.

[DEX, 1996] DEX. Dicționarul explicativ a limbii Române. I.Coteanu, Luiza Seche, Mircea Seche. Univers enciclopedic. Bucuresti, 1996.

[MULTEXT Specifications] MULTEXT. Specifications and Proposed Standards Available at http://www.lpl.univ-aix.fr/projects/multext/MUL3.html

[MULTEXT-East Specifications] MULTEXT-East lexical specifications. Concede Edition. Available at http://nl.ijs.si/ME/V2/msd/

[Ratnaparkhi, 1998] Adwait Ratnaparkhi. 1998. Maximum Entropy Models for Language Ambiguity Resolution. *A dissertation presented at the Faculties of the University of Pensylvania.*

[Shieber, 1986] Stuart M. Shieber. 1986. An introduction to unification-based approaches to grammar. *Center for the study of language and information.*

[Steina and Nagao, 1997] Steina, J. and Nagao, M. 1997. Corpus Based PP Attachment Ambiguity Resolution with a Semantic Dictionary. In Zhou, J. and Church, K., editors, *Proceedings of the Fifth Workshop on Very Large Corpora*, pages 66-80, Beijing and Hong Kong.

[Tomaz Erjavec et al., 1998] Tomaz Erjavec, Nancy Ide, Dan Tufis. 1998. Development and Assessment of Common Lexical Specifications for Six Central and Eastern European Languages. *ALLC-ACH '98 Conference*, Debrecen, Hungary, July.

[Tufis, 2000] Dan Tufis. 2000. Using a large set of EAGLES-compilant morpho-syntactic descriptors as a tagset for probabilistic tagging. *In Proceedings of the second International conference on Language Resources and Evaluation,* pages 1105-1112, Athens, Greece, May.