Knowledge-based Coreference Resolution for Hungarian

Márton Miháltz
MorphoLogic
Orbánhegyi út 5, Budapest, 1126 Hungary
E-mail: mihaltz@morphologic.hu

Abstract

We present a knowledge-based coreference resolution system for noun phrases in Hungarian texts. The system is used as a module in an automated psychological text processing project. Our system uses rules that rely on knowledge from the morphological, syntactic and semantic output of a deep parser and semantic relations form the Hungarian WordNet ontology. We also use rules that rely on Binding Theory, research results in Hungarian psycholinguistics, current research on proper name coreference identification and our own heuristics. We describe the constraints-and-preferences algorithm in detail that attempts to find coreference information for proper names, common nouns, pronouns and zero pronouns in texts. We present evaluation results for our system on a corpus manually annotated with coreference relations. Precision of the resolution of various coreference types reaches up to 80%, while overall recall is 63%. We also present an investigation of the various error types our system produced, along with an analysis of the results.

1. Introduction

In this paper, we describe a knowledge-based NP-coreference resolution system, one that identifies antecedents in a text that refer to the same entities in the real world as the anaphoric noun phrases for Hungarian.

Our system deals with the following types of anaphoric phenomena (pairs of corefering noun phrases set in bold in the examples):

Type	Example (Hungarian, English)
Repetition	Tegnap találkoztam **egy ismerősömmel**. **Az ismerősöm** nagyon sietett. *I met **an acquaintance** today. **My acquaintance** was in a hurry.*
Proper Name Variant	**Kovács Jakab, az ABC Kft. igazgatója** tegnap sajtótájékoztatót tartott. Az eseményen **Kovács úr** bejelentette az új termékeket. ***Jakab Kovács, chairman of ABC Ltd.** held a press conference today. **Mr. Kovács** announced the new products.*
Synonym	Tamás kapott **egy biciklit**. Én is láttam **a kerékpárt**. *Tamás got **a new bicycle**. I saw **the bike**, too.*
Hypernym	Bejött **egy kutya**. **Az állat** fáradtnak tűnt. ***A dog** just came inside. **The animal** seemed tired.*
Pronoun	Beszéltem **Julival**. Megadtam **neki** a számodat. *I talked to **Juli**. I gave **her** your phone number.*
Zero Pronoun	**Viktor** ismeri **Ferit**, de (**ő**) nem kedveli (**őt**) túlságosan. ***Viktor$_1$** knows **Feri$_2$**, but **he$_1$** doesn't like **him$_2$** very much.*

Table 1: Examples for the types of coreference we attempt to resolve.

The case of zero pronouns is a phenomenon in Hungarian when the pronominal arguments of the main verb are phonologically empty – the suffixes on the verb carry enough information about the number and person of the arguments –, but otherwise require the same treatment as regular, phonologically not empty personal pronouns.

At present, we do not deal with cataphora (when the antecedent is preceded by the anaphora). We also don't handle subcomponents of complex noun phrases (possessive structures, coordination, deverbal nouns with their arguments etc.), only simple, maximal NPs corresponding to the arguments of the main verb. At this stage, our system only handles personal pronouns but no other types of pronouns.

In the following section, we describe the design principles and the general coreference resolution algorithm, followed by detailed description of the system for the various types of anaphora. In section 3, we describe the results of evaluating the performance of our system against a small, manually tagged corpus. In the last section, we discuss possibilities to improve the system.

2. Description of the Coreference Resolution System

In recent work in the field of coreference resolution (CR), data-driven, machine learning-based approached have gained ground over traditional knowledge-based systems (Ng, 2005). However, such an approach requires an extensive number of hand-labeled training examples, which is not available at present for Hungarian. For this reason, we had to commit ourselves to a rule-based approach in the design of our CR system.

Our system relies on several sources of knowledge. The most important knowledge source is the morphological, syntactic and semantic information available from the output of the MetaMorpho MT system's deep parser (Prószéky et al 2004). Rules based on Binding Theory in Hungarian syntax (Kenesei, 1992) and the results of psycholinguistic research on Hungarian sentence understanding (Pléh, 1998) operate on these structures. Rules based on semantic relationships are based on information in the Hungarian WordNet (HuWN, Miháltz et al, 2008). For the matching of proper name variants, we employ character-based heuristics, similar to some of those described by Uryupina (2004).

The MetaMorpho parser identifies paragraph, sentence and token boundaries, clauses, maximal noun and verb phrases, and provides morphological, grammatical and semantic information for these. After this preprocessing, our system processes each anaphoric NP in the document from left to right and tries to identify the corefering antecedent that is closest to it. This means that pronouns and zero pronouns can be the antecedents of pronouns or zero pronouns as well.

Before coreference resolution, we try to identify as much as possible from the NPs that are formally anaphoric but are likely to refer to entities outside the document. This is done in order to exclude NPs that would only introduce noise to the CR process (Varasdi 2005).

Coreference resolution for a given NP in the input document is based on satisfying constraints, in order to eliminate as much as possible from the antecedent candidates and evaluating preferences in order the select the most likely candidate (Mitkov, 1999). The algorithm for generating the list of antecedent candidates, filtering the list and finally selecting the winning candidate is specific to the type of the anaphoric NP (proper name, definite common noun, pronoun/zero pronoun) and is described in detail below.

2.1 Proper names

For proper names, the list of antecedent candidates consist of all the proper names prior to the anaphor in the entire document. At present, we do not apply any kind of filtering to these candidates. The most likely antecedent candidate is the one having smallest Minimum Edit Distance (MED) with the anaphor. Both antecedent and anaphor are normalized before the string matching: determiners are removed from the beginnings of the names, and the head word is lemmatized. The rule selects an antecedent only in case the MED for the closest candidate falls below a preset threshold. This way, the system is not forced to select one from the available candidates.

2.2 Common nouns

For common nouns with a definite article, we first try to exclude mentions that refer to unique objects inferable from common world knowledge (e.g. "the president of the United States"). At present, we do this by searching a predefined list of NPs. Antecedent candidates are the proper names and common nouns in the preceding part of the paragraph of the anaphor, up to the VP containing it (Binding Theory excludes candidates dominated by the main verb in the anaphor's VP.) Selecting the antecedent is done by identifying the closest candidate that has the same head, or the closest synonym or hypernym/hyponym. Synonymity is checked via Hungarian WordNet: if there is a synset that contains both anaphor and candidate they are considered synonyms. We use the Leacock-Chodorow similarity measure (Leacock and Chodorow, 1998) in order to measure semantic relatedness via the hypernym/hyponym paths connecting the anaphor and the candidate (lexical forms of the heads are used.) The closest candidate that falls below a preset threshold is considered as the winning antecedent, but only if no identical or synonymous candidate was found. In the evaluation experiment described in section 3, the threshold was configured to accept candidates available in WN not further than 2 relation "steps" away.

2.3 Pronouns, zero pronouns

We only deal with personal pronouns, with the exception of *az* ("that") demonstrative pronoun in subject position and not referring to a subordinate relative clause (explanation follows). We exclude 1st and 2nd person pronouns (referring to discourse entities outside the text). The antecedent candidates are collected from the 2 sentences before the anaphor's sentence (if they exists) plus the clauses prior to the clause containing the anaphor in its sentence. All kinds of NPs in this scope are considered.

The antecedent candidates are filtered by checking person, number and 2 semantic features specified by the parser: *+/-animate* and *+/-human.* The latter two can have underspecified values (in case of zero pronouns and lexically ambiguous nouns), these are compatible with all other values. The filtering process also excludes candidates that have already been identified as antecedents of other NPs in the current clause (in accordance with Binding Theory.)

If there is more than one pronominal anaphor in the current clause, the system processes them in obliqueness order (the subject first, then the (direct) object, then the other valence arguments, and finally the modifiers.) This allows the simple identification of antecedents by ruling out already bound candidates (see above.) We also perform CR for common nouns and proper names before resolving pronouns within a sentence. This is done in order to further help resolution of pronouns, which are the most difficult, but also the most frequently occurring type of anaphor, according to our preliminary observations.

Identifying the antecedent of the pronoun or zero pronoun that is the subject in its VP follows research on Hungarian psycholinguistics (Pléh, 1998.) The heuristic first assumes parallel grammatical functions across sentences, where the subject is preserved from the previous clause/sentence. This is overridden by the presence of the demonstrative pronoun *az* in subject position, which indicates change of subject. In case there are more than one non-subject NPs in the prior clause, the antecedent is selected using the obliqueness hierarchy and by checking distance from the anaphor (NPs closer to the end of the sentence are preferred). Pléh describes other indicators of subject change (such as semantic preference of arguments by predicates), but at the present stage, we do not deal with these phenomena. Resolution of pronouns and zero pronouns with grammatical roles other than subject are based on the obliqueness hierarchy and closeness to the anaphor.

3. Evaluation

At the present stage of our work, we have carried out a preliminary evaluation in order to assess the performance of our CR system. We have compiled a small corpus from excerpts from history textbooks, one of the focus areas of the psycholinguistic text processing project that utilizes our CR system. The texts in the corpus were processed with MetaMorpho to annotate structural and grammatical boundaries. The mentions identified by the parser were manually annotated by their closest antecedents in the texts. We note that the annotated corpus is not complete in

that it does not cover all instances of coreference present in the text, only the ones possible to mark between NPs correctly identified by the parser.

Table 2 shows the statistics for our small evaluation corpus:

Texts	10
Paragraphs	31
Sentences	99
NPs	488
NPs annotated with coreference (all types)	132
NPs annotated with coreference (types handled by system)	81

Table 2: Parameters of the evaluation corpus.

We used 16 different types of NP-coreference for the manual annotation, which had 132 occurrences in the corpus. 6 of these types are handled by our current CR system, which gives 81 annotated NPs for testing. Table 3 shows the distribution of the various types of NP coreference annotated in the corpus.

Coreference Type	Number of occurrences
Personal pronoun	**47**
Possessive NP	21
Repeated NP	**15**
Proper name variant	**14**
Demonstrative pronoun	8
Frame	7
"that"-clause	6
Hypernym	**3**
Relative pronoun for relative clause	3
Wh-pronoun in relative clause	2
Synonym	**2**
Apposition	1
Copula	1
Hyponym	**1**
Meronym	1
Holonym	0
Total:	*132*

Table 3: Type and number of occurrences of coreference annotated in the evaluation corpus. Coreference types handled by our system are set in boldface.

We performed coreference resolution for the texts in the corpus with our system and compared the results with the manual annotation. We calculated precision (the ratio of correctly resolved NPs to the number of NPs tagged by the system) and recall (the ratio of correctly resolved NPs to the number NPs manually annotated) for each type of coreference we presently handle (Table 4.) We regarded automatically tagged references correct that were not identical to the annotated reference for the NP but belonged to the same coreference chain, ie. referred to the same entity.

After a first look at the results, we were able to confirm that the system performs fairly well (precision 71-80%, recall 61-83%) for the most frequent types of anaphora currently handled in the corpus (proper name variant matching, repeated forms of common nouns and pronouns, zero pronouns.) On the other hand, the performance of the synonym and hypernym heuristics was poor, but since the evaluation corpus contained only a small number of such instances, this figure might not reflect realistic evaluation.

We also conducted an examination of the various error types produced by the system. Each automatically assigned coreference link was examined, and assigned to one the four categories:

- OK: coreference link produced by system is identical to manual annotation (correct).
- OK_equ: coreference link produced by system is not identical to manual annotation, but refers to same entity (in the coreference chain), so it was regarded correct.
- KO_parser: coreference link assigned by system is different from manual annotation (ie. erroneous); the error is due to erroneous syntactic parsing in the input (if the parser would have provided correct results, the automatic coreference assignment would have been correct.)
- KO_cr: erroneous result; the antecedent was present in the text and the parsing was correct; the mistake was due to the CR algorithm.

As it can be seen from Table 5, about half of all the mistakes committed by the CR system are results of errors in parsing, such as incorrectly tagged noun phrases, zero anaphors etc. Having perfectly parsed input would increase overall precision to 75%, pronoun/zero pronoun resolution precision to 91%. This tells us that our method is rather sensitive to parsing accuracy in the input.

Finding not exactly matching, but referentially equivalent antecedents is a phenomenon only observed in the case of pronouns/zero pronouns, as the pronominal CR algorithm mainly relies on tracking of the discourse. Tracing back to the beginning of coreference chains in order to label the first mention of each entity as antecedent would result in lower precision, due both to parsing errors and CR errors.

Coref. Type	OK	OK_equ	KO_parser	KO_cr
Pronoun	19	6	7	3
Repeated	13	0	4	1
Prop. name	12	0	0	3
Hypernym	0	0	0	2
Synonym	1	0	0	3
Hyponym	0	0	0	0
Total:	45	6	11	12

Table 5: Categorization of evaluation results.

We have also experimented with a second round of evaluation in order to compare our results to a previous work on Hungarian anaphora resolution by Lejtovicz

(2006), which uses an implementation of Centering Theory (Brennan, Friedman and Pollard, 1987). At the present stage, the coreference type covered by both Lejtovicz's and our system is zero pronouns in subject positions. So far we have selected 3 news articles from the Szeged Treebank (Csendes et al, 2005), which has accurate, manually created syntactic annotation. There were 15 anaphora occurrences in the selected articles, which were first manually labeled with their antecedents, then compared to the results of running Lejtovicz's and our system. Both systems had very low coverage (4 and 3 anaphors attempted), of which 3 were correct (75% and 100% precision). We will continue to annotate coreference in selected texts from the Szeged Treebank in order to be able to compare our systems on the basis of more data that is not dependent on the output of a specific parser.

	NPs manually	NPs tagged				
Coreference type	annotated	Total	Correct	Precision	Recall	F-measure
Proper name	14	15	12	80.00%	85.71%	82.76%
Pronoun	46	35	25	71.43%	54.35%	61.73%
Repeated	15	18	13	72.22%	86.67%	78.79%
Synonym	2	4	1	25.00%	50.00%	33.33%
Hypernym	4	2	0	0.00%	0.00%	0.00%
Total/Average:	*81*	*74*	*45*	*68.92%*	*62.96%*	*65.81%*

Table 4: Precision, recall and f-measure for the different coreference types, measured on the manually annotated corpus.

4. Present and Future Work

After an examination of examples from the corpus, we have come up with a number of ideas which could be used to improve the performance of our current system.

To further extend the criteria for matching proper name variants, the MED function could be complemented by accounting for the named entity types (e.g. person, company, country, city etc.) We plan to use MetaMorpho's built-in NE-classifier for this purpose.

A relatively frequent phenomenon is when a singular noun phrase that designates a group (e.g. a company name) is referred to by a pronoun in plural form (e.g. "they"). In these cases, filtering candidates by number should be overridden by semantic type information (ie. group nouns).

Recognizing idiomatic complements of verb phrases – noun phrases in the valence frame that are not referential, eg. "give a *hand* to somebody" – would help to further eliminate non-referential NPs from attempting coreference resolution. We plan to use MetaMorpho's output for this purpose, since its grammar contains this information.

The second component in noun-noun endocentric compounds in Hungarian is often the hypernym of the first component, such as in English a *grand piano* is a kind of a *piano*. Using a morphological analyzer these can be effectively utilized to discover semantic information (Mihăltz, 2003). Using this method would provide a novel way to experiment with treating hypernym coreferences.

To extend the coverage of coreference resolution, we plan to incorporate rules to acknowledge coreference between subject and predicate of nominal predicate clauses (e.g. "He is the winner") and the subjects and objects of predicates like *is called*, *his/her name is* etc.

We would also like to work on handling possessive references, which were the second most frequent type of coreference in our test corpus (see Table 3.) We also plan to create rules to handle appositions and relative pronouns in the near future to handle further varieties of coreferential phenomena.

5. References

Brennan, Susan, Friedman, Marilyn W. and Pollard, Carl J.: A centering approach to pronouns. In Proceedings of the 25th Meeting of the Association for Computational Linguistics (ACL '87), (1987), pp.155-162.

Csendes D., Csirik J., Gyimóthy T., Kocsor A.: The Szeged Treebank. In Proc. of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005), Karlovy Vary, Czech Republic pp. 123-131 (2005).

Kenesei, István: Az alárendelt mondatok szerkezete. (The Structure of Complex Sentences.) In: Kiefer Ferenc (ed.): Strukturális Magyar Nyelvtan, vol. I., Mondattan. Akadémiai Kiadó, Budapest (1992)

Leacock, C., M. Chodorow: Combining Local Context and WordNet Similarity for Word Sense Identification. In C. Fellbaum (ed.): WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA (1998), pp. 265–285

Lejtovicz, Katalin, Kardkovács, Zsolt: Anaphora Resolution in Hungarian Texts. In Proceedings of The Fourth Conference on Hungarian Computational Linguistics (MSZNY 2006), Szeged (2006), pp. 362-363.

Mihăltz, Márton: Constructing a Hungarian Ontology using Automatically Acquired Semantic Information. In Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5), Tilburg, The Netherlands (2003), pp. 475-478.

Mihăltz, M., Cs. Hatvani, J. Kuti, Gy. Szarvas, J. Csirik, G. Prószéky, T. Váradi: Methods and Results of the Hungarian Wordnet Project. In: Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, (2008), pp. 311–321.

Mitkov, Ruslan: Anaphora Resolution: The State of The Art. Working Paper, University of Wolverhampton

(1999)
Ng, Vincent: Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. Proceeding of the 43rd Annual Meeting of the Associaction for Computational Linguistics (2005)
Pléh, Csaba: Mondatközi viszonyok feldolgozása: az anafora megértése a magyarban. (Processing Intrasentential Relationships: The Understanding of Anaphora in Hungarian) In: Pléh Csaba: Mondatmegértés a magyar nyelvben. Osiris Kiadó, Budapest (1998)
Prószéky, Gábor; László Tihanyi; Gábor Ugray: Moose: a robust high-performance parser and generator. Proceedings of the 9th Workshop of the European Association for Machine Translation, Foundation for International Studies, La Valletta, Malta, pp. 138–142 (2004)
Uryupina, Olga: Evaluating Name-Matching for Coreference Resolution. In Proceedings of the 4th International Conference on Language Resources and Evaluation (2004)
Varasdi, Károly: Koreferenciák feloldása (Coreference resolution). Manuscript (2005)