Anaphora resolution experiment with CG rules

Tiina Puolakainen
University of Tartu, Estonia
tiina.puolakainen@ut.ee

Abstract

This article describes an ongoing work - an experiment on anaphora resolution in Estonian newspaper texts using Constraint Grammar (CG) rules. The personal, demonstrative and relative pronouns are chosen for resolution at the first stage of the research. As morphological information and syntactic relations play an important role in identifying anaphoric relations, syntactically analyzed text is used as an input for the CG rules.

1 Introduction

The experiment is performed on a subtask of coreference resolution task, considering only pronouns, in particular personal, demonstrative and relative pronouns, some of the latter can serve as interrogative pronouns as well in particular sentences. Reference is exophoric if the referred entity is brought into the textual space and pronoun refers to those objects of the real world that the speakers can see or imagine. Reference is endophoric if the referred item has been mentioned earlier in the text or there is an intention to do so immediately. Endophoric reference can be called anaphoric if the antecedent of the pronoun has occurred previously in the text or cataphoric if it will follow the pronoun (Pajusalu, 1996).

The antecedent of a pronoun in at text is usually a noun or a noun phrase or a noun in an adpositional phrase, but it can be also an adjective in case of adjectival pronoun or it may refer to an entire clause in case pronoun refers to a situation expressed by the whole clause. Finally the antecedent of a pronoun can be also another pronoun - then it can refer further to it's antecedent and so compose a chain of pronouns hopefully ending by a meaningful reference.

The latest version of the CG formalism (Karlsson et al., 1995) - VISL CG3 framework (Bick, 2000) is chosen for the needs of the experiment because this framework is very flexible and precise at the same time. It enables to write as general rules as possible and as specific as needed - in that sense it can be considered a programming language that is proved to be suitable and successful for many language processing tasks.

Preliminary work consisted of extracting of synonyms and synosets with hyponym relation for some task-specific notions as speech-verbs and animate-inanimate nouns from the Estonian Wordnet (Vider and Orav, 2005). The extracted sets however could not be used entirely because they included much more submeanings of the desired concepts than could be affordable. The resulting sets were filtered manually and then included into CG grammar as a word lists used by the CG rules.

2 Anaphora resolution

The procedure takes a syntactically analyzed Estonian text from the Estonian Dependency Treebank (Muischnek et al., 2014). Sentences have been analyzed by syntactic parser (Muischnek, Müürisep and Puolakainen, 2014) and then manually checked and corrected by a linguist. It is possible to use just automatically analyzed sentences but in this case more errors occur due to remaining morphological and syntactical ambiguities and also errors made by the syntactical analyzer.

First the pronouns are identified in the text by adding a special tag to the word analysis. Then

the relation identification rules are applied for each pronoun.

CG grammar for identification of coreferences consist of three types of rules according to the type of pronoun. Personal pronouns are also differentiated by the person (1st, 2nd, 3rd) and all pronouns by the number (singular, plural). Demonstrative pronouns referring to more general situation expressed by a clause as well as relative-interrogative pronouns in interrogative sentences receive a reference relation to a predicate of the corresponding clause. Special attention has to be paid to the direct speech and as a special case - a dialog or interview-style text. The additional rules have to be added to enable to deal with different styles of presenting such text, currently only the style that is encountered in a training corpus is maintained by the rules.

Taking advantage of known syntactical functions, usually subject and object are preferred by the rules if choosing the relation candidate as subject and object as core arguments are more likely to carry the main theme of the story. But in specific situations other considerations are used by the rules as for example the cases of complements of speech-verbs indicating the theme of the talk while searching for antecedent of demonstrative pronoun.

An animate-inanimate sets were worth of using them but should be further examined and subcategorized for fitting their aim more precisely. The difficulty in using these sets arises due to the quite free grammatical choice of corresponding pronoun, especially in a case of inanimate concept that can be referred with personal pronoun. The most frequent example for that is a concept of some organisation, that can be referred both as inanimate and animate substance, in the latter meaning the people of that organisation. In this case even number 'agreement' between pronoun and its antecedent may not hold. Pronoun can be in plural and it's antecedent in singular also for other generalizations. According to Pajusalu (1996), prototypically *tema* 'he/she' refers to animate and *see* 'it' to inanimate entity, but the real use of pronoun system is much more complicated. For example, quite common is to use *tema* to refer to more concrete entity and *see* for more abstract one. But it depends very strongly on particular topic, for example music can be considered concrete when talking about particular composition and an abstract in general, it depends also on the formality of the situation. From two relatively similar entities in the animate-inanimate and abstract-concrete axes first is usually referred as *tema* and second as *see*. There are definitely many more nuances of different usage of pronoun system and all kinds of metaphors also cause this kind of difficulties for recognizing coreferences.

3 Evaluation

For evaluating the rules a newspaper text of 2080 words was taken from Estonian Dependency Treebank. The test benchmark consisted of four stories from different newspapers and was not especially selected to contain very complicated coreference relations. 150 CG rules for anaphora resolution were applied to the syntactically analyzed text and then manually checked for correctness of marked reference. The text contained 101 pronouns, of which 79 received correct reference relation that means a 78% recall. Applying the same procedure to the same but automatically analyzed text increases the amount of errors from 22 to 28 giving recall of 72%. Table 1 shows results of the evaluation according to the type of the pronouns, numbers in parentheses indicate the corresponding values then applied on automatically analyzed text. Most difficulties are imposed by demonstrative pronouns as they have also the widest spectrum of entities they can refer to, including exophoric reference where any suggestion of an antecedent in the text become wrong. The easiest is resolution of reference of relative pronouns, if only they are correctly distinguished from the interrogative usage.

	Demonstrative	Personal	Relative	Total
Correct no	20 (18)	30 (27)	29 (28)	79 (73)
Correct %	59 (53)	81 (73)	97 (93)	78 (72)
Incorrect no	14 (16)	7 (10)	1 (2)	22 (28)
Incorrect %	41 (47)	19 (27)	3 (7)	22 (28)
Total	34	37	30	101

Table 1. Evaluation results of anaphora resolution.

An example of a sentence (1) with successfully recognized reference relations both in a case of a treebank sentence and an automatically syntactically analyzed sentence:

(1) *Digitaalühenduse (1) saavad need (→3) kliendid (2), kes (→2) seda (→1) soovivad (3).*

Digital-connection-sg.gen (1) receive that-pl.nom (→3) clients (2) who (→2) this-sg.prt (→1) want (3).

'The digital connection (1) can be received by whose (→3) clients (2), who (→2) want (→3) that (→1).'

'That' points to 'digital connection', 'who' to 'clients' and 'whose' refer to the subordinate clause 'who want that' specifying the kind of 'whose' clients.

In the following sentence (2) two relations were recognized correctly and third remained unrelated with both types of input as rules could not determine the exact referred situation in the previous context:

(2) *Ma (→2 sentences ahead) (1) ei tea, kui hästi see (→none) mul (→1) õnnestub.*

I (→2 sentences ahead) (1) not know how well that (→none) I-sg.ade (→1) succeeds.

'I (→2 sentences ahead) (1) do not know, how well it (→none) (for me (→1)) succeeds.'

Situations with multiple suitable candidates and/or no sufficiently good candidate are difficult to handle. As the rules are applied individually, one by one, each of them has to be quite careful for their decision. In some circumstances no rule in the set of rules formulated so far can decide the correct antecedent. Mostly this indicates 'gaps' that can be filled by formulating new rules.

(3) *Puutüved (1) on miljoneid aastaid (1') vastu pidanud tänu sellele (→2), et neid (→1/1') ümbritses (2) kiht liiva (3), mis (→3) võis olla tekkinud mõnest tugevast liivatormist.*

Tree-trunks (1) are million-pl.prt year-pl.prt (1') against hold-ppp thanks this-sg.all (→2) that this-pl.prt (→1/1') enclosed (2) coat sand-sg.prt (3) that (→3) might have evolved some-sg.ela strong-sg.ela sandstorm-sg.ela.

'The treetrunks (1) have survived for millions of years (1') due to the fact (→2) that they (→1/1') have been enclosed (2) by a coat of sand (3) that (→3) might have evolved as a result of a strong sandstorm.'

In sentence (3) all three relations are correctly recognized using a treebank sentence as a input, but makes one mistake using automatically syntactically analyzed sentence: 'they' is found to refer to 'millions of years' instead of 'The trunks' due to remained morphological ambiguity in the sentence.

One particular source of errors in automatically syntactically analyzed text is remaining ambiguity between *tema* 'he/she' and *see* 'it' pronouns that have some homonymous forms or in the worse case the error made during disambiguation of these forms. There is no good solution so far for this kind of errors. On the one hand disambiguation module needs the information of referenced entity to make right decision choosing between two pronouns, but on the other hand coreference resolution module expects that the right decision is already done by disambiguation module. Also wrong case and subsequently incorrectly chosen syntactical function can cause additional errors in automatic mode.

One more source of errors is the distance between pronoun and it's antecedent. Choosing far distances could allow to find also far relations but at the same time it would introduce more errors there unrelated entities would be marked as related. The CG framework allows to choose a window size for the working distance of the rules and after some experiments a window of 7 sentences was chosen as an optimal working load. This caused one error in test set where the correct antecedent was in the 9th sentence ahead.

4 Using synonym relations

The hypothesis for this experiment was that antecedent of a demonstrative pronoun can be situated in a very similar context, namely containing synonym words. In order to check this hypothesis all synonyms in the text at a distance of maximally 7 sentences from each other were found using Estonian Wordnet hierarchy of synosets and up to 3 layers of hypo- and hyperonym relations. After that rules were applied if encountered pronoun and potential antecedent in the context of synonyms.

An example (4) of a successful application of such rule: *hõõgus* 'smouldered' is found to be a synonym of *põleks* 'could burn' and due to that the subject pronoun *see* 'it' in one sentence was related with it's antecedent - subject in the sentence with synonym predicate.

(4) ... *turvas* (SUBJ) hõõgus (SYN) eredalt isegi päevavalguses ...

... *peat* (SUBJ) smouldered (SYN) brightly even in day light ...

... (some sentences)

Vabalt põleks (SYN) *see* (SUBJ) veel vähemalt 4 nädalat.

It (SUBJ) could burn (SYN)at least 4 weeks further.

In one more successful example (5) pronoun *neist* 'form them' was correctly related to an antecedent NP, but, in fact, verbs that are found to be synonyms in these sentences are not used in their synonym submeanings – *käima* as 'go' and *tulema* in it's modal sense 'have to', not it's main sense 'to come'.

(5) Jooksu *intensiivsus ja maht* käivad (SYN) käsikäes.

The *intensity and amount* of running go (SYN) hand in hand.

Üht *neist* suurendades tuleb (SYN) teist vähendada.

Increasing one *of them* (you) have to (SYN) reduce the other.

Overall, this experiment didn't bring an improvement, but did even worse, because found synonym and close hypo- and hyperonym relations appeared to be mostly too loose or general and didn't account for different submeanings and caused too many errors.

5 Conclusions

The results of the small experiment on usual newspaper text are good for the beginning - in 72-78% of cases pronouns receive correct reference relation. The experiment shows the main difficulties and limitations of approach encountered during anaphora resolution and gives the directions for further work, including the research of wider range of coreference types besides pronouns. The main aim is to use other sources of semantic relations and methods of distributive semantics besides wordnet that can help to solve the task. The many-to-many synonym relations of wordnet synosets cannot be efficiently realized by the means of CG framework and need to be handled in a separate preprocessing step.

References

Eckhard Bick. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus University Press, Aarhus, UK.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, Arto Anttila. 1995. *Constraint Grammar: A Language-Independent System for Parsing Running Text.* Natural Language Processing, No 4. Mouton de Gruyter, Berlin and New York.

Kadri Muischnek, Kaili Müürisep and Tiina Puolakainen. 2014. Dependency Parsing of Estonian: Statistical and Rule-based Approaches. In *Proceedings of the Sixth International Conference: Human Language Technologies – The Baltic Perspective.* Amsterdam: IOS Press, (Frontiers in Artifical Intelligence and Appications; 268), pages 111-118.

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, Dage Särg. 2014. Estonian Dependency Treebank and its annotation scheme. In *Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13).* Tübingen, pages 285-291.

Renate Pajusalu. 1996. Pronoun Systems of Common Estonian and Estonian Dialects in Contrastive Perspective. *Estonian Typological Studies I.* Tartu, pages 145-164.

Kadri Vider and Heili Orav. 2005. Estonian wordnet and Lexicography. *Symposium on Lexicography XI. Proceedings of the Eleventh International Symposium on Lexicography.* Lexicographica, Series Maior 115. Max Niemeyer Verlag, Tübingen, pages 549-555.

Abbreviations used in glosses

ade adessive case
all allative case
gen genitive case
ela elative case
nom nominative case
pl plural
prt partitive case
ppp past participle
sg singular