TOWARDS BETTER PARSING OF SPOKEN ESTONIAN

Kaili Müürisep, Helen Nigol
University of Tartu, Estonia

Abstract

This paper reports on our ongoing work to develop a parser for spoken Estonian using written language parser as a basis. In the first stage of the project we presuppose that the input is morphologically analysed and unambiguous and we focus only on syntactic phenomena of spoken language. As the spontaneous speech is rich of so called ungrammatical constructions and disfluencies a special attention has been paid to smooth such utterances before the classical parsing. The parser gained the recall 97.7% and precision 90.4% in the analysis of the benchmark corpus.

Keywords: Spoken language, disfluency detection, parsing, Estonian.

1. Introduction

Parsing of spontaneous speech is a serious challenge: spoken language has often different vocabulary, it is hard to determine where the sentence starts from and where it ends due to the lack of capitalized letters and punctuation marks. Spontaneous speech is also rich of ungrammatical constructions like unfinished, elliptical or overlapping utterances, parenthesis, mispronunciations, truncated words, filled pauses, repetitions, false starts, self-repairs etc. The common term of these phenomena is called as disfluency (Eklund 2004). One type of disfluency that has proven particularly problematic for parsing is speech repairs: when a speaker amends what he is saying mid-sentence.

We use the rule-based parser of written Estonian (Müürisep 2001) for syntactic annotation of Corpus of Spoken Estonian. The parser is based on Constraint Grammar (CG) framework (Karlsson et al. 1995). The parser gives a shallow surface oriented analysis to a sentence, in which every word is annotated with the tag corresponding to its syntactic function (in addition to morphological description). The head and modifiers are not directly linked, only the tag of modifiers indicates the direction where the head may be found. The CG parser consists of two modules: morphological disambiguator and syntactic parser. In this paper, we presume that the input (transcribed speech) is already morphologically unambiguous and the word forms have been normalized according to their orthographic forms.

In the following sections we will describe our corpora for developing parser and the main traits of spoken language which complicate automatic syntactic analysis.

2. Syntactically analysed corpus of spoken Estonian

The corpus of spoken Estonian (1 065 000 words, 1703 transcripts) contains 100 000 part-of-speech-tagged and manually disambiguated words (Hennoste et al. 2000). Our goal is to provide syntactic annotation to that part of the corpus. The texts were normalized (vaguely articulated or colloquial words have the description of the corresponding word form in the written language) and provided with some transcription annotation (longer pauses, falling or rising intonation).

For the assessment of the work of the parser, the benchmark corpus of 6700 words was compiled and analysed manually by a single expert.

We used separate corpus of 8400 words for training the parser (i.e., generating or modifying rules) and in addition the special corpus of disfluencies (Nigol 2007) which was annotated according to principles of the *Disfluency annotation stylebook for the Switchboard corpus* (Meteer et al. 1995). Based on this corpus of disfluencies two syntactically annotated corpora were created. The first corpus was parsed in its original form; the second was parsed after its normalization.

These corpora contain both longer narrative dialogues and shorter dialogues where turns alternate swiftly.

3. Parsing transcribed speech

Although spontaneous transcribed speech is rich of unfinished, elliptical or overlapping utterances, parenthesis, mispronunciations, truncated words, hesitated and filled pauses, repetitions, false starts, self-repairs and other spoken language specific phenomena, the first decision we had to start with was what is the sentence in spoken language. In the case of the spoken language, the segmentation of the speech into sentences is a complex task. In the written language, the sentence is considered as one syntactic unit or syntactic window which defines the scope of context used for parsing. The sentence of the written language has punctuation mark-up added by the author of the sentence. In the spoken language, the falling and rising intonations may be used as the delimiters of the sentences but this may not be precise enough - the intonation may often fall inside the syntactic unit also. This was the reason why we decided to consider a turn in the dialogue as a syntactic window and treat the units separated by full stops as coordinated clauses. (The style of the determination of the sentence boundary depends on the type of the text, and the usage of full stops as sentence delimiters is justified in the case of monologues.)

Even using dialogue turns as sentences may not be satisfactory enough since the sentence can extend the borders of dialogue turns both in the case of overlaps or the speakers construct one sentence together (see example 1).

(1) H: Ja aadress on Liivi
'H: And the address is Liivi Street'

K: kaks
'K: two'

The adaption of grammar rules of written language for spoken language was easier task than we supposed. The main stress was on reformulating inner-clause boundary rules. The clause boundaries in written text are determined by the rules based on conjunctions, punctuation marks and verbs. These clause boundary detection rules have been thoroughly revised since the meaning and usage of punctuation marks have been changed. The new rules take account the intonation mark-up, pauses, special particles and of course conjunctions and verbs. In spite of the that, the erroneous clause boundaries remain the main source of errors: more than the third of errors in training corpus were caused by faulty or missing clause boundaries.

The syntactic constraints needed only minimal modification which consider the less precise detection of clause boundaries and different vocabulary.

The recall of the parser was 97.3% and precision 89.2% (Müürisep et al. 2006). The corresponding numbers of the parser for written language were 98.5% and 87.5% respectively.

The better results may be explained by the small size of the corpus (2500 words), but after the enlargement of the benchmark corpus to 6700 words the results remain approximately same (recall 97.6.%, precision 90.2%) (Müürisep et al. 2007). The actual reason should be that the utterances are shorter and simpler than sentences of written language and there are more particles with unambiguous syntactic function in spoken language than in written texts and the statistics is word based. In the other hand, these results show that the Constraint Grammar formalism is really suitable for parsing any unrestricted running text. The similar experiment is described by Eckhard Bick (1998) for Portuguese, adapting Constraint Grammar based tagger/parser for written Portuguese as a tool for annotating Brazilian urban speech corpus.

The analysis of errors showed that the main sources of errors were erroneous detection of inner clause boundaries and the defective constructions or disfluencies.

Parsing both types of corpora of disfluencies demonstrated that parser gains 0.5-3% better results if the nonfluent parts of the utterances have been removed before the automatic processing (see Table 1).

Table 1. The impact of disfluencies in the results of the parser

Type of disfluency	Utterances	Recall	Precision
Repairs	original	94.4	84.6
	normalized	96.2	87.3
Repetitions	original	98.2	90.7
	normalized	98.6	91.8
False starts	original	97.4	90.0
	normalized	98.9	93.8

3. Detection of disfluencies

As seen from the Table 1, the repetitions are the easiest type of disfluencies. In most cases repetitions do not cause any harm to the parsing since mostly the words which are repeated belong to the conjunctions or particles (for example: *and and and then he made a mistake*). The repetition of nouns is critical since if a subject or an object is repeated then the principle of uniqueness has been violated.

We use a grammar external script for tagging simple repetitions of a single word (*miks miks miks peab* ...- /why why why one should .../). The repetitions of verb *be* and numerals may occur in the normal sentence, so we had to consider the part-of-speech tags.

Also, it is possible to repeat the same word in different cases (see 2). Some of these repetitions are normal in written texts also (e.g, *samm sammult* /step by step/), but the others signal to the occurrence of a self-repair.

(2)	noh	erinevatel	päevadel	on	võimalik	siis
	noh	different	days	is	possible	then

mägi	mäge	valida
hill-NOM	hill-PART	to choose

'so it is possible to choose a hill in different days'

One of the signals about self-repairs are the truncated words (e.g. *nor-* instead of *normal*). All these words are tagged with a special tag - @REP (repair or repetition). Also the patterns *word break- word* and *word break- või* were detected and annotated.

These words and phrases are removed (commented out) from the input of the parser and added back to the output. This simple technique gave good results in the processing of utterances with repetitions and helped a little in the case of self-repairs (see Table 2).

The false starts have been eliminated by clause boundary rules of the grammar, separating these into independent chunks. False starts are easy to recognize if they contain a verb. The parser do not try to mark up false starts with special tag, the unfinished sentence is analyzed as it is possible using the present grammatical information.

Table 2. Results with tagged repetitions and self repairs

	Recall	Precision
Repairs	94.6	85.1
Repetitions	98.6	92.1
False starts	98.1	91.1
Test corpus	97.7	90.4

As the disfluencies are such a heterogeneous class of linguistic events, maximum results may only be achieved through combining different methods. Previous studies in English (see (Eklund, 2004) for brief overview) have showed that only lexical patterns are not enough for detecting disfluencies. More information, especially at the acoustic-prosodic level (e.g. intonation, duration, pauses), is no doubt needed to reliably detect repairs and false starts.

4. Shortcomings and open questions of our approach

Transcriptions of spontaneous speech are hard texts for syntactic analysis even for linguists. It is very difficult to decide which is a correct annotation of a word in the unfinished or grammatically incorrect sentence. We use special tag - @T (unknown syntactic function) for clear situations. One should decide which syntactic tag to use if the human annotator may guess the syntactic function but the grammatical information is incomplete for the parser (see example 3).

(3)	Ta	võttis	selle	punase	noh	tead küll
	He	took	this	red	well	you know

Should the words *this* and *red* be premodifying attributes, object or unknown? Our annotation is still somehow inconsistent although we try to avoid the label for unknown syntactic function if possible.

The similar problems arise during annotation of repairs. The example 4 demonstrates the case where reparator may be analysed as postmodifying apposition.

(4) mind uvitaksid Tallinnas asuvad kirjastused + raamatukirjastused
I am intrested in publishers + book publishers in Tallinn

Is it correct to use disfluency annotation in the same level as syntactic annotation? Should we use special mark-up for false starts? What to do with unfinished sentences? And what about grammatically incorrect sentences which have never been repaired by the speaker? Should we distinguish semantic and syntactic self-repairs (compare examples 5 and 6)?

(5) ma sain homseks või tändab esmaspäevaks piletid
I got the tickets for tomorrow I mean for Monday

(6)	mh	poola	poole	öheksaks	tulevad	tuleb	see	meister
	mh	polish	half	nine	come	comes	this	master

'mh, this master will come for half past eight'

We need to enhance the annotation scheme of our corpora with more exact disfluency information, similar to the works of (Bies et al. 2006).

5. Conclusions and plans for future

The rapid progress in the field of speech recognition has increased the interest in the topic of disfluency detection but the developed methods are mostly data-driven which presuppose huge annotated corpora. Our approach to both parsing and disfluency detection is rule-based which fits our resources and goal best.

The experiment for improving the efficiency of the parser demonstrated that the grammar written originally for written unrestricted text is suitable for parsing spoken language but one should pay a special attention to the automatic identification of disfluencies.

The parser gained the recall 97.7% and precision 90.4% in the analysis of the benchmark corpus. The parser performs relatively well in the analysis of repetitions (the recall and precision gained in the automatic analysis are approximately the same as the results achieved from the corpus where the repetitions had been removed manually). The detection of false starts and self-repairs needs further development.

Our future plans are to polish the annotation scheme, finish the shallow syntactic annotation of the corpus and the further goal is to transform the corpus to treebank.

The next challenge would be also the adaptation of morphological disambiguator to spoken language and even to make experiments to apply the parser to the output of speech recognition system.

References

Bick, Eckhard 1998. Tagging Speech Data - Constraint Grammar Analysis of Spoken Portuguese, *Proceedings of the 17th Scandinavian Conference of Linguistics* Odense.

Bies, A., Strassel, S., Lee, H., Maeda, K., Kulick, S., Liu, Y., Harper, M.,Lease, M. 2006. Linguistic resources for speech parsing, in *Fifth International Conference on Language Resources and Evaluation (LREC'06*), Genoa, Italy

Eklund, Robert 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Linköping Studies in Science and Technology, Dissertation No. 882, Department of Computer and Information Science, Linköping University, Sweden.

Hennoste, Tiit; Lindström, Liina; Rääbis, Andriela; Toomet, Piret; Vellerind, Riina 2000. Tartu University Corpus of Spoken Estonian. – T. Seilenthal, A. Nurk, T. Palo (eds.). *Congressus Nonus Internationalis Fenno-Ugristarum*. Pars IV. Dissertationes sectionum: Linguistica I, Tartu. 345-351.

Karlsson, Fred; Anttila, Arto; Heikkilä, Juha; Voutilainen, Atro 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*. Mouton de Gruyter.

Meteer, M.; Taylor, A.; MacIntyre, R.; Iver, R. 1995. *Dysfluency annotation stylebook for the Switchboard corpus*. Distributed by LDC.

Müürisep, Kaili 2001. Parsing Estonian with Constraint Grammar. *Proceedings of NODALIDA'01*. Uppsala.

Müürisep, Kaili; Uibo, Heli 2006. Shallow Parsing of Spoken Estonian Using Constraint Grammar. In: Henrichsen, P. J.; Skadhauge, P.R. (eds). *Treebanking for Discourse and Speech. Proceed. of NODALIDA 2005 Special Session on Treebanks for Spoken Language and Discourse.* Copenhagen Studies in Language 32. Samfundslitteratur. 105-118

Müürisep, Kaili; Nigol, Helen 2007. Disfluency Detection and Parsing of Transcribed Speech of Estonian. *Proceedings of Human Language Technologies as a Challenge for Computer Science and Linguistics*. 3rd Language & Technology Conference (ed. Zygmunt Vetulani). Poznan, Poland. Fundacja Uniwersitetu im. A. Mickiewicza. 483-487.

Nigol, Helen 2007. Parsing Manually Detected and Normalized Disfluencies in Spoken Estonian. *Proceedings of NODALIDA 2007*. Tartu.

KAILI MÜÜRISEP is a senior researcher of Institute of Computer Science, University of Tartu. She received her PhD in computer science from University of Tartu in 2000. Her research topics are focussed on automatic syntactic analysis of Estonian. E-mail: kaili.muurisep@ut.ee.

HELEN NIGOL is a PhD student at Institute of General and Estonian Linguistics. She received her MA. in general linguistics from University of Tartu in 2006. She is interested in syntactic phenomena of disfluencies in spoken language. E-mail: helen.nigol@ut.ee.