Automatic Sentence Simplification for Subtitling in Dutch and English

Walter Daelemans and Anja Höthker and Erik Tjong Kim Sang

CNTS - University of Antwerp
{walter.daelemans, anja.hoethker, erik.tjongkimsang}@ua.ac.be
Universiteitsplein 1, Wilrijk, Belgium

Abstract

We describe ongoing work on sentence summarization in the European MUSA project and the Flemish ATraNoS project. Both projects aim at automatic generation of TV subtitles for hearing-impaired people. This involves speech recognition, a topic which is not covered in this paper, and summarizing sentences in such a way that they fit in the available space for subtitles. The target language is equal to the source language: Dutch in ATraNoS and English in MUSA. A separate part of MUSA deals with translating the English subtitles to French and Greek. We compare two methods for monolingual sentence length reduction: one based on learning sentence reduction from a parallel corpus and one based on hand-crafted deletion rules.

1. Introduction

This paper addresses work in progress on the language engineering aspects involved in automating the production of subtitles for television programmes on the basis of text (written transcripts) and ultimately on the basis of speech recognition output. The natural language processing task involved in this problem is to simplify sentences by removing lexical material or by paraphrasing parts of them. The result should be a reduction of the sentence's length in characters down to a value computed dynamically from constraints on the speed with which subtitles can be shown and the size of the region on the screen they may occupy. As for other NLP tasks, both statistical (machine learning) and linguistic knowledge-based techniques can be considered for this problem. Given that we have available a considerable amount of data in the form of transcripts of programmes with their associated subtitles, a machine learning approach can at least be investigated.

In the remainder of this paper, we describe the project context in which this task is investigated and the data and other resources that have been developed. We characterise the task as being related to but different from document summarization and describe related work. We go on to compare the accuracy of a machine learning approach to a knowledge-based approach. The former is based on inducing sentence reduction generalizations from an aligned monolingual corpus using machine learning algorithms, the latter on hand-crafted deletion rules using a robust shallow parser for our Dutch and English data. We point at the problems with evaluation for this task, and provide a first interpretation of the results.

2. ATraNoS and MUSA resources for sentence simplification

The context of this research is a common subproblem (sentence simplification) in the Flemish project ATraNoS[1] and the European project MUSA[2]. Both projects aim at automatic generation of TV subtitles (for hearing-impaired people or for multilingual access when combined with translation). This involves speech recognition, sentence simplification, and, in MUSA, machine translation. The target language for the sentence simplification task is equal to the source language: Dutch in ATraNoS and English in MUSA The machine learning part of the latter project deals with translating the English subtitles to French and Greek.

For the sentence simplification task, the following resources were constructed or adapted (both for English and Dutch).

- *Parallel corpora of programme transcripts and subtitles.* The Dutch material consists of news broadcasts obtained from the public Belgian TV company VRT and the Dutch TV company NOS as well as a small section with episodes of the Flemish VRT soap Thuis. The English material contains documentaries and talk shows provided by the British BBC World Service. The VRT news sections used in this paper contain about 430,000 words in the subtitle part and the English BBC material about 400,000 words.

- *Alignment software.* The corpora have been aligned on sentence level with an alignment method based on lexicalised similarities (Vandeghinste and Tjong Kim Sang, 2004). The alignment software obtained precision and recall figures of 91% for linking sentences in the VRT corpus. Data that present problems to the alignment algorithm are sentence duplicate sentences in either the subtitles or the transcripts, subtitles that contain more words than the related transcript and sentences with spelling variations. In order to improve the quality of the corpus, all alignments have been checked manually.

- *Shallow parser.* A shallow parsing approach based on memory-based learning was adopted for linguistically analyzing these corpora (Daelemans et al., 1999; Buchholz et al., 1999; Van den Bosch and Daelemans, 1999; Buchholz, 2002; Tjong Kim Sang, 2002) resulting in the assignment to all sentences in the corpus of lemmas and part-of-speech tags to the words, and syntactic phrase labels to related adjacent words. The shallow parser modules also perform basic relation

[1] Partners: ELIS Ghent, CCL Leuven, ESAT Leuven, and CNTS Antwerp

[2] Partners: ILSP Athens, ESAT Leuven, SYSTRAN Paris, LCC Athens, BBC London, and CNTS Antwerp

finding (identification of verbs and their subjects, objects and other relations). Additionally, proper names in the text were identified and classified. These linguistic annotations will be used both in the machine learning and in the rule-based approach.

3. Approaches to sentence simplification

Although at a superficial level the task of subtitle generation seems similar to the problem of automatic summarization (see for overviews Mani and Maybury 1999, and Mani 2001)) there are important differences. Summarization is usually taken to mean the production of a shorter version of an orginal document or set of documents by keeping the most informative parts of the original either by selecting sentences on the basis of measures of salience or by template-based information extraction. The problem of automatic subtitling is easier in that there is no problem of sentence selection and in general the compression needed is relatively limited (15% on average for BBC documentaries). On the other hand, the problem is more difficult because existing, correct, sentences from the original cannot be reused making it hard to produce shorter sentences that are both syntactically and semantically coherent. An additional problem is that the amount of compression needed is computed dynamically and should be adhered to as closely as possible, and that processing should be on-line, so that the complete document cannot be processed before compression.

We are not aware of previous work on sentence simplification for automatic subtitling. Similar approaches have been developed for other applications, however. Grefenstette (1998) applies shallow parsing and simplification rules to the problem of *telegraphic text reduction*, with as goal the development of an audio scanner for the blind or for people using their sight for other tasks like driving. Another related application area is the shortening of text to fit the screen of mobile devices (Corston-Oliver, 2001; Euler, 2002). The latter uses a list of statistically relevant words and syntactic reduction constraints to simplify sentences.

Hori (2002) uses dynamic programming to create abstracts from transcribed speech through word extraction combining different scores (word relevance, linguistic likelihood, confidence measure and word concatenation probabilities). In Caroll et al. (1998) and Canning and Tait (1999), the application area is the production of text understandable by people suffering from aphasia. In the PSET project, text simplification rules guided by the properties of patients with aphasia are applied to linguistically analysed newspaper text. Of course, also in document summarization, the generation of compact sentences capturing the most salient information is a useful subprocess mimicking how people construct summaries, and delivering better solutions than simply concatenating extracted sentences (even when problems of anaphora resolution and cohesion can be solved). For that reason, sentence level compression has received attention there as well (Chandrasekar et al., 1996; Knight and Marcu, 2002; Jing and McKeown, 1999).

In this paper we will compare two approaches to sentence simplification. The first is a machine learning approach in which a simplification model is learned from parallel corpora with TV programme transcripts and the associated subtitles. The second approach is a knowledge-based approach which relies on hand-crafted phrase deletion rules. We are interested in learning the strengths and weaknesses of both methods and in finding out whether a combination of the two might outperform the best individual approach.

4. Machine learning approach

For the machine learning experiments, we have represented the summarization process as a word transformation task: words in the transcribed text can be copied, deleted or replaced. Copying a word is the most frequent action present in the parallel corpus. Word insertions have been ignored. The performance of the summarization process has been measured with precision, recall and F$_{\beta=1}$ rates for word deletions and word replacements. The latter only were correct in those cases that a word in the transcript was replaced by the same word as in the subtitle. Apart from these three evaluation rates we have also registered compression rates: the length of the subtitle in characters divided by the length of the transcript sentence.

The sentences in the corpora were aligned on word level by linking identical words and word-paraphrase pairs that appeared in a paraphrase dictionary. After this, we have selected interesting sentence pairs: those in which the transcript was different from the subtitle but which shared at least half of the words of the longest sentence in the pair. This resulted in a Dutch corpus with 12,535 sentences (156,701 words) and an English corpus with 6,164 sentences (108,015 words). From these pairs we kept 90% as training material for the machine learner. The remaining 10% of the data was used as test material (also for the rule-based method described below).

A memory-based learner (Daelemans et al., 2002) was applied to the training data. It was fed with words, lemmas, part-of-speech tags, chunk tags, relation tags and proper name tags. Apart from the focus word we also included information regarding a context of two words to the left and right. The learner thus had 30 features to its disposal.

A feature selection process was performed with bidirectional hill-climbing (Caruana and Freitag, 1994) in order to determine an optimal set of features. For Dutch, the selection process chose word features, lemma features, part-of-speech features and chunk features but neither name features nor relation features. The machine learner obtains an F$_{\beta=1}$ rate of 24.3 for Dutch (92.3% compression rate; the target was 81.3%). The performance rates for English were F$_{\beta=1}$=15.5 with a character compression rate of 96.4% where 87.1% was the target. The English learner selected word features, lemma features and chunk features, but neither part-of-speech features, name features nor relation features. A baseline learner which predicted the most frequent output tag for each word had obtained F$_{\beta=1}$=1.6 for Dutch and F$_{\beta=1}$=0.3 for English (see Table 1).

The machine learning approach did not perform as well as we had expected. The performances were low and, most importantly, the approach frequently made nonsensical errors, like removing sentence subjects or deleting a part of a multi-word unit. We have encoded enough information in the features to prevent such errors. Unfortunately, with the present data sets, an optimal performance measured in F$_{\beta=1}$ rates allows the usage of only a limited number of features thus limiting the awareness of the system to all-but local information. We believe that in order for the machine learning approach to perform better, we would need a much larger amount of data, something which is beyond the scope of the projects because of the costs of the required manual checks.

5. Rule-based approach

In order to get a better understanding of the problem of sentence simplification, we decided to manually compile phrase deletion rules for the two languages. The deletion rules have access to the same syntactic information as the machine learner. However, in order to avoid the rules being tuned to the two parallel corpora, we have based them on external non-parallel data. Our goal was to perform phrase deletion in two steps: first selecting all phrases that are more or less redundant, and second choosing some of the phrases for deletion in such a way that the required compression rate is met.

The phrase deletion rules include among others rules for removing adverbs, adjectives, first names, prepositional phrases, phrases between commas or brackets, relative clauses, numbers and time phrases. Some of the rules used for the English data are listed below:

- Noun phrases: we keep the head word of noun phrases and mark the preceding words for deletion unless omitting them would make the sentence ungrammatical (determiners, pronoun) or alter the meaning of the sentence (comparative, superlative).

- Prepositional phrases: from a grammatical point of view it is safe to remove all prepositional phrases.

- Adjectives: all adjectives are suggested for deletion. Special care has to be taken to delete neighbouring conjunctions if necessary.

- Adverbs: adverbs can always be omitted unless vital for the meaning (never, not,...). Another exception is comparing context (run as fast as you can).

- Sentence initial conjunctions and interjections can be removed without loss of grammaticality or information.

After all eligible rules have been applied we do a final check to ensure that we did not keep sentence parts that actually belong to a fragment that was a candidate for deletion (for example possessive endings).

We have employed two different selection methods for choosing phrases for deletion. Candidate phrases for deletions in the Dutch data were ordered by length (shortest first) and sentence position (first phrases towards the end of the sentence). Candidate deletion phrases were removed by selecting the phrases containing the smallest number of words first while using the position of the phrase in the sentence as a tie-breaker. The deletion process continued until the required compression rate was obtained. The deletion rules obtain an F$_{\beta=1}$ rate of 26.2 for Dutch (the compression rate was 74.3%).

Dutch	Precision	Recall	F$_{\beta=1}$	CR
Baseline	59.6%	0.8%	1.6	99.6%
Learner	42.0%	17.1%	24.3	92.3%
Rules	26.4%	26.1%	26.2	74.3%
Combined	26.8%	28.0%	27.4	74.8%

English	Precision	Recall	F$_{\beta=1}$	CR
Baseline	60.0%	0.1%	0.3	100.0%
Learner	33.1%	10.1%	15.5	96.4%
Rules	25.1%	18.3%	21.2	83.5%
Combined	25.3%	20.3%	22.5	83.6%

Table 1: Performances of the machine learner, the rule-based approach and a combination of the two on the two data sets measured in the percentage of remaining characters (compression rate: CR) and precision, recall and F$_{\beta=1}$ obtained on word deletions and word replacements. The baseline system predicts the most frequent action for each word. The target character compression rates are 81.3% for Dutch and 87.1% for English.

In the English process, candidate phrases for deletions were sorted by surprise values: the log-likelihood of the frequencies of the words in the phrases as extracted from a large corpus. Additionally, the English reduction process contains a preprocessing step in which we looked up common phrases from a table of paraphrases. This provides a reliable and accurate way of achieving sentence compression, at a low cost. The paraphrases were extracted semi-automatically from available transcripts and their hand-made subtitles. We ran the experiments with and without the preprocessing step. It turned out that the preprocessing step did not have an effect on the overall F$_{\beta=1}$ rates: in both cases we obtained 21.2 (Table 1). Both the Dutch and English rules performed significantly better than the machine learner with respect to F$_{\beta=1}$ rates ($p<0.05$ according to bootstrap resampling) and with respect to obtaining the required compression rates although the latter is not very surprising given that the learner did not have explicit access to the required compression rates.

Since the rules performed only phrase deletions and not phrase replacements, we have combined the machine learner with the rules by selecting all word replacements predicted by the learner and relying on the decisions made by the rules for the other words. The combined approach outperformed the deletion rules for both languages although the differences were not significant ($p\sim0.1$ for Dutch and $p\sim0.2$ for English, see Table 1).

6. Concluding remarks and future work

In this paper we have described a machine learner approach to sentence simplification in which a system learned the simplification task from parallel corpora of TV programme transcripts and the associated subtitles. This approach did not work very well, most likely because our training corpora are too small. Unfortunately it is unlikely that we will obtain access to large high-quality parallel corpora because compiling them requires a lot of manual labor.

The next approach we have chosen for sentence simplification for subtitling contains three steps: 1. replacing phrases by shorter ones (paraphrasing), 2. finding candidates for phrase deletion, and 3. selecting phrases for deletion. The second step was performed by sets of handcrafted deletion rules. For the third step we evaluated two approaches, both of which were satisfactory although a more detailed comparison of the two remains to be done. The phrase deletion part of this approach already outperformed the machine learner and the complete approach with the learner selecting the paraphrases did even better (Table 1).

As in summarization research at large, evaluation of automatic subtitling is problematic (see Mani et al. 2002 for a recent overview of summarization evaluation). As is also the case for NLP tasks such as translation and prosody generation, there are often several correct solutions, and comparison to a gold standard as we did in this paper is limited in that sense. The most reliable evaluation would be to have humans evaluate the output of the system in terms of semantic and syntactic well-formedness. However, this approach is not feasible in the development phase of a system, where the effect of design decisions, learning algorithm parameters, information sources etc. has to be judged without overfitting on a single test set. We are currently investigating the BLEU methodology developed in the context of Machine Translation (Papineni et al., 2002; Hori et al., 2003) for system development, and human evaluation for the final system.

Acknowledgements

Anja Höthker is supported by the European Union's Fifth Framework Programme as a researcher in the MUSA project. Erik Tjong Kim Sang is funded by IWT STWW as a researcher in the ATraNoS project.

7. References

Buchholz, Sabine, 2002. *Memory-Based Grammatical Relation Finding*. Ph.D. thesis, Tilburg University.

Buchholz, Sabine, Jorn Veenstra, and Walter Daelemans, 1999. Cascaded Grammatical Relation Assignment. In *Proceedings of EMNLP/VLC-99*. pages 239–246.

Canning, Yvonne and John Tait, 1999. Syntactic Simplification of Newspaper Text for Aphasic Readers. In *ACM SIGIR'99 Workshop on Customised Information Delivery*. Berkeley, CA, USA, pages 6–11.

Carroll, J., G. Minnen, Y. Canning, S. Devlin, and J. Tait, 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*. pages 7–10.

Caruana, Rich and Dayne Freitag, 1994. Greedy Attribute Selection. In *Proceedings of the Eleventh International Conference on Machine Learning*. Morgan Kaufman, pages 28–36.

Chandrasekar, R., Christine Doran, and B. Srinivas, 1996. Motivations and Methods for Text Simplification. In *COLING-96*. pages 1041–1044.

Corston-Oliver, Simon, 2001. Text compaction for display on very small screens. In *Proceedings of the Workshop on Automatic Summarization, NAACL*.

Daelemans, Walter, Sabine Buchholz, and Jorn Veenstra, 1999. Memory-based shallow parsing. In *Proceedings of CoNNL-99 / Computational Natural Language Learning*. pages 53–60.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, 2002. *TiMBL: Tilburg Memory Based Learner, version 4.3, Reference Guide*.

Euler, Timm, 2002. Tailoring Text Using Topic Words: Selection and Compression. In IEEE Computer Society Press (ed.), *Proceedings of 3rd International Workshop on Natural Language and Information Systems (NLIS)*. pages 215–219.

Grefenstette, Gregory, 1998. Producing Intelligent Telegraphic Text Reduction to provide an Audio Scanning Service for the Blind. In *Intelligent Text Summarization, AAAI Spring Symposium Series*. pages 111 – 117.

Hori, Chiori, 2002. *A Study on Statistical Methods for Automatic Speech Summarization*. Ph.D. thesis, Graduate School of Information Science and Engineering Tokyo.

Hori, Chiori, Takaaki Hori, and Sadaoki Furui, 2003. Evaluation Methods for Automatic Speech Summarization. In *Proceedings of Eurospeech 2003*. Geneva, Switzerland.

Jing, Hongyan and Kathleen McKeown, 1999. The Decomposition of Human-Written Summary Sentences. In *Research and Development in Information Retrieval*. pages 129–136.

Knight, Kevin and Daniel Marcu, 2002. Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. In *Artificial Intelligence*, volume 139(1). pages 91–107.

Mani, I., T. Firmin, D. House, G. Klein, B. Sundheim, and L Hirschman, 2002. The TIPSTER SUMMAC Text Summarization Evaluation. In *Natural Language Engineering*, volume 8, 1. Cambridge University Press.

Mani, Inderjeet, 2001. *Automatic Summarization*. John Benjamins.

Mani, Inderjeet and Mark T. Maybury (eds.), 1999. *Advances in Automatic Speech Summarization*. Cambride, Massachusetts: MIT press.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2002. BLEU: a method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. pages 311–318.

Tjong Kim Sang, Erik, 2002. Memory-based Shallow Parsing. In *Journal of Machine Learning Research*, volume 2. pages 559–594.

Van den Bosch, Antal and Walter Daelemans, 1999. Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99*. pages 285–292.

Vandeghinste, Vincent and Erik Tjong Kim Sang, 2004. Using a Parallel Transcript/Subtitle Corpus for Sentence Compression. In *Proceedings of LREC 2004*. Lisbon, Portugal.