Condensing sentences for subtitle generation

Prokopis Prokopidis*, Vassia Karra†, Aggeliki Papagianopoulou†, Stelios Piperidis*

*ILSP-Athena RC
Artemidos 6 & Epidavrou, GR 15125, Athens, Greece
prokopis@ilsp.gr, spip@ilsp.gr

†University of Athens
vassiakarra@yahoo.gr, angeliki4@gmail.com

Abstract

Text condensation aims at shortening the length of an utterance without losing essential textual information. In this paper, we report on the implementation and preliminary evaluation of a sentence condensation tool for Greek using a manually constructed table of 450 lexical paraphrases, and a set of rules that delete syntactic subtrees that carry minor semantic information. Evaluation on two sentence sets show promising results regarding grammaticality and semantic acceptability of compressed versions.

1. Introduction

Text condensation aims at shortening the length of an utterance without losing essential textual information. Text compression tools have been used as building blocks in text simplification, automatic summarization, headline generation, subtitle generation, and information extraction applications. In this paper, we report on the implementation and preliminary evaluation of a text condensation tool for Greek. Our architecture consists of two main stages: a) word substitution using a manually constructed paraphrase table and b) elimination of elements that carry minor semantic information via a set of rules operating on sentences represented as dependency trees.

2. Related work

(Knight and Marcu, 2000) discuss a noisy-channel and a decision-tree approach to the problem of sentence compression, using a set of 1067 English sentences and their compressed versions for training and testing. Human evaluation on 32 sentence pairs showed average scores of 4.3 and 3.5 for grammaticality and importance (whether most important information is retained in the compression) on a scale from 1 to 5 for the decision-based model. Working on the same set of sentences, (Riezler et al., 2003) used a maximum entropy model to select the most probable compression from reduced f-structures generated by an LFG grammar. In a headline generation scenario, (Dorr et al., 2003) use linguistically motivated rules for iterative shortening of peripheral constituents in parsed sentences. For machine-assisted subtitling in Catalan, (Bouayad-Agha et al., 2006) present a set of compression strategies that include removal of repetions, number to digit transformations, and deletion of non-important linguistic units. The work presented in our paper is based on the methodology for English and Dutch subtitle generation described in (Daelemans et al., 2004).

3. Architecture

In our processing architecture, input text is first channeled through a set of tools for shallow linguistic analysis, including a tokenizer, a sentence splitter, a POS tagger and a lemmatizer (Papageorgiou et al., 2002). In the next processing stage a paraphrase module matches the words of the input text with those contained in a paraphrase table (detailed in §4) and when possible, substitutes them for their semantic equivalents. Following this stage, the text is further compressed via a set of deletion rules described in §5.

In the evaluation experiments described in this paper, annotated input is compressed until no more paraphrases and deletion rules are applicable. In a variant of this scenario, our system for automatic generator of subtitle drafts sets a desired compression rate, expressed in requirements for the deletion of n words and/or m characters. In this case, paraphrase application and deletion rules are applied until the compression rate is reached. However, in the setting of multilingual subtitle generation, where subtitles have to be automatically translated in another language, the condensation subsystem may produce versions that satisfy larger compression requirements, taking into account the possibility of constraint violation in the output of the translation process.

Compressed sentences pass through a final processing stage, during which they are converted into visually correct subtitles (Karamitroglou, 1998). More precisely, using the results of the automatic syntactic analysis detailed in §5, we segment sentences into more than one subtitles at the highest syntactic nodes possible. We also try to proportionally distribute words in two-line subtitles, and we do not allow more than one sentences in the same subtitle.

4. Paraphrase Resource

The paraphrase module currently implemented aims at lexical paraphrasing, i.e. word(s) substitution without any kind of syntactic transformation. The module uses a manually compiled table of 450 paraphrases between 503 lemmata. The table is encoded as an XML document. During paraphrase matching in our text condensation scenario, only paraphrases resulting into shorter versions of the source word are of course considered.

We used a thesaurus of synonyms and antonyms (Iordanidou, 2006) to construct an initial seed of paraphrase lemmata. Paraphrases that were too domain- or register-

```
<Paraphrase id="p400" stag="AjBaFeSgGe"
source="αναπάντεχης" target="αιφνίδιας"/>

<Paraphrase id="p6501" source="θιασώτες"
stag="NoCmMaPlAc" target="οπαδούς" />
<Paraphrase id="p6503" source="θιασώτες"
stag="NoCmMaPlNm" target="οπαδοί" />

<Paraphrase id="p523" source="έχω την
εντύπωση" stag="VbMnIdPr01SgXxIpAvXx"
target="νομίζω""/>
```

Figure 1: Examples of paraphrases between types sharing the same morphological features

specific were omitted. We then evaluated this seed against a large corpus of morphosyntactically annotated Greek texts (Hatzigeorgiu et al., 2000), checking for paraphrase interchangeability and applicability in different linguistic contexts. When all morphological variants of each lemma were automatically generated, we came up with a set of 9860 paraphrases between types sharing similar morphological features. For example, the two adjectives αναπάντεχης and αιφνίδιας (unexpected, sudden) in the first example of Figure 1 share the same morphological values (*Fe*male gender, *S*ingular number, *Ge*nitive case). Since input text is automatically annotated for the same features, this information guides the paraphrase module into making correct substitutions for homographic source types that may correspond to more that one target types. Thus, in the case of θιασώτες (followers, *Nom*inative—*Ac*cusative), the module will choose between synonym types οπαδοί kai οπαδούς based on the case of the source word. The third example in Figure 1 presents a paraphrase pair involving the multi-word έχω την εντύπωση (I have the impression) and νομίζω (I think).

The paraphrase table contains both unidirectional and bidirectional paraphrases. A pair of paraphrases was stored as unidirectional if it consisted of lemmas that we qualitatively judged to be not mutually interchangeable in most linguistic contexts. Bidirectional paraphrases are pairs that were judged to be semantically equivalent and interchangeable in most syntactic contexts. For instance, the lemmas πρωτάκουστος (never heard before) and πρωτοφανής (never seen before) are used to show that something is *novel* or *unprecedented*, and can therefore substitute for each other in most contexts without affecting sentence meaning.

5. Deletion rules

After paraphrase matching, (and if, in our subtitle generation application, compression requirements have not been met) we use a set of rules in order to delete sentence elements that carry minor semantic information, in an effort to ensure that the meaning and grammatical correctness of the output remain relatively intact. The rules take as input sentences represented as dependency trees. For this type of syntactic representation, we exploit the MaltParser platform (Nivre et al., 2004), via which we have trained a memory-based dependency parser for Greek. The parser was trained on the 70K words of the Greek Dependency Treebank, which comprises data annotated at several linguistic levels (Prokopidis et al., 2005). The dependency label set comprises 25 main relations.

Each deletion rule traverses the nodes of the dependency tree, checking whether specific morphosyntactic constraints apply for the node currently examined. When the constraints match, the node and the subtree that is headed by this node are marked as *deletables*. Constraints may focus on the node's (or children or parent nodes') dependency relations, their POS tag, etc. As an example, we can examine *delAdjs*, one of the rules most often applied. This rule, a simplified version of which is shown in Figure 2, marks adjectives which a) are not the heads of other nodes, and b) are not labeled *Pnom*, i.e they are not headed by a copula verb.

Subtrees marked to be deleted are ranked according to their *relevance*, which is estimated as in (Daelemans et al., 2004) on the basis of the log-likelihood of the frequencies of the words in the subtrees' nodes as observed in a Greek newspaper corpus of 70M words. Using this information, deletion of the least significant subtrees, which is expected not to seriously affect sentence meaning, precedes elimination of more important subtrees.

6. Evaluation

In a first evaluation experiment, we randomly chose 100 automatically parsed sentences (TestSetA) and examined the condensation achieved by the combined use of paraphrase lookup and application of deletion rules, as a) the average number of characters deleted and b) the average compression rate achieved. After initial results with loss of large pieces of important information due to removal of long subtrees, we decided to specify a maximum limit for deletion, not allowing a rule to apply if more than 5 nodes were to be deleted.

In the example of Figure 3, we show one sentence from TestSetA, its translation, and three actions of the text condensation tool. The first action is the application of the paraphrase παραξενεύει → ξενίζει (both 3rd person, singular types of verbs that can be translated as *surprise*) , with a gain of 4 characters (c_del=4). The other two actions remove two adjectives from the input, resulting in a gain of 1 word each (w_del=1). The adjective σκληρός (hard) is deleted after όλα (all) since its relevance is higher.

As shown in Table 1, a rather low rate of textual compression (2.78 chars per sentence) is achieved via paraphrasing. On the other hand, deletion rules reduce the original length of a sentence by 18.8 characters. As the average sentence length in our sample was 117 characters, average compression rates of 2.37% and 16.06% were reached by paraphrasing and deletion rules, respectively. This is of course due to the fact that deletion rules are more general than a rather limited, manually compiled, paraphrase table. Nevertheless, compression requirements in our subtitle generation system are often quite modest (requesting, for example, deletion of a few characters), and thus simple word replacement with shorter paraphrases can prove useful.

```
sub delAdjs {
  my $node = shift;
  if ($node->children()) {
    # recursively apply rule if there are children nodes
    foreach my $child ($node->children()) {
      delAdjs($child);
    }
  # An adjective not headed by a copula verb
  } elsif (($node->getAttribute("tag") =~ /^Aj/)
       && ($node->getAttribute("afun") ne /Pnom/)) {
    # mark as deletable
      setDeletable($node);
  }
}
```

Figure 2: Example of a deletion rule

```
Orig: Αυτό δεν πρέπει να μας παραξενεύει καθώς το πακέτο καταλαμβάνει περισσότερα από 10MB χώρου
στον σκληρό δίσκο όταν αποσυμπιεστούν όλα τα αρχεία των δισκετών εγκατάστασης .
(This should not surprise us, as the package occupies more than 10MB in the hard
disk when all the installation disk files are decompressed.)

actn_348_1: Paraphrase (w_del=0, c_del=4): παραξενεύει -> ξενίζει
actn_348_2: delAdjs (rel=7.45 - w_del=1, c_del=3): όλα (all)
actn_348_3: delAdjs (rel=10.47 - w_del=1, c_del=6): σκληρό (hard)

Result: Αυτό δεν πρέπει να μας ξενίζει καθώς το πακέτο καταλαμβάνει περισσότερα από 10MB χώρου
στον δίσκο όταν αποσυμπιεστούν τα αρχεία των δισκετών εγκατάστασης.
(This should not surprise us, as the package occupies more than 10MB in the disk
when the installation disk files are decompressed.)
```

Figure 3: Reducing sentence length via three condensation actions

ACTIONS	DEL.CHRS	COMPR
paraphrases	2.78	2.37%
deletion rules	18.8	16.06%

Table 1: Condensation results for TestSetA expressed in deleted characters and compression rate achieved

In a second evaluation experiment, we used a set of 100 sentences (TestSetB) that were manually annotated at the level of syntax. Sentences in TestSetB were relatively short, with an average sentence length of 39.13 tokens. We asked two human judges to evaluate the grammaticality and the semantic acceptability of the iteratively compressed versions of the original sentences, using a scale from 1 to 5. The judges examined the result of 375 condensation actions, summarized in Table 2.

The three most frequent actions involve deletions of adjectives (delAdjs), adverbs (delAdvs), and preposition-headed adverbials (delPPs). The first two do not reduce drastically the original sentence length. Nevertheless, their application does not lead to ungrammatical sentences. Moreover, they were considered by the evaluators as meaning preserving, since average semantic acceptability scores for each type of rule were 4.2 and 4.5, respectively. Other rules involving deletion of, among other structures, relative or adverbial clauses are more aggressive but, due to the 5 node limit, less frequently applied.

ACTIONS	FREQ	DEL.CHRS
delAdjs	205	8.89
delAdvs	87	6.04
delPPs	44	17.64
other rules	12	19.4
all rules	348	9.56
paraphrases	27	1.92
all actions	375	9.01

Table 2: Application frequency of condensation actions on TestSetB, and average gain in deleted characters for each type of action.

Table 3 summarizes evaluation judgments for final output versions, i.e. compressed sentences after all possible actions have been applied. As expected, there is a trade-off between compression rate and quality of the output, with average scores dropping, for example, from 4.48 for grammaticality of compressions up to 10%, to 3.50% accord-

ing to the first evaluator. Since semantic acceptability is a more vague term compared to grammaticality, it is perhaps reasonable to observe a worse interannotator agreement in the case of the former. We will investigate setting more strict criteria for categorizing compressed versions in future work.

	GR1	GR2	SEM1	SEM2
All	4.10	4.76	3.15	4.14
$< 10\%$	4.48	4.89	3.55	4.32
$10 - 20\%$	4.14	4.83	2.97	4.14
$> 20\%$	3.50	4.45	3.00	3.90

Table 3: Average judgements from two human evaluators on a 1-5 scale, as far as grammaticality (GR1, GR2) and semantic acceptability are concerned (SEM1, SEM2). Judgements for <10%, 10-20%, and >20% compressed sentences are also shown.

Let us examine an example where grammaticality and semantic acceptability conflict. In the clause όταν ολοκληρωθούν οι γλωσσικές αποδόσεις (gloss = *when complete_3rdPersonPlural the language renderings*, translation = *when translations are complete*), one evaluator gave a 5 score regarding the effect the deletion of the adjective γλωσσικές had on the grammaticality of the sentence, and a 3 regarding semantic acceptability; 5 and 2 scores were given by the other annotator. We should notice that although this isolated clause becomes unintelligible once the adjective is deleted, its context (the discussion about a document that, in order to be read by members of the European Parliament, typically has to be translated first) may in this case help the reader disambiguate the sense of the remaining noun.

7. Conclusions and Future Work

We presented a viable architecture for condensation of Greek sentences via paraphrasing and deletion rules. Our preliminary evaluation experiments show promising results as far as grammaticality of the compressed versions is concerned. Semantic acceptabilily is a more vague, and more context-dependent, notion, and we intend to focus on incorporating more clues from whole documents on whether particular sentence segments are important or not. In order to improve the coverage of our approach, we plan to augment our paraphrase table with paraphrases semi-automatically extracted from parallel corpora consisting of program transcripts and their corresponding hand-crafted subtitles.
The tool described in this paper is currently being used in the context of a machine-assisted multilingual subtitling environment for Greek television broadcasts. In another context, and in order to use our module for text simplification purposes, we will introduce rules that transform complex syntactic structures into simpler ones.

Acknowledgements

We would like to thank three anonymous reviewers for useful suggestions and comments. Work described in this paper was supported by research grants Sub4All and Interreg.

8. References

N. Bouayad-Agha, A. Gil, O. Valentin, and V. Pascual. 2006. A Sentence Compression Module for Machine-Assisted Subtitling. In *Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 490–501, Mexico City, Mexico.

D. Daelemans, A. Höthker, and E. Tjong. 2004. Automatic Sentence Simplification for Subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048, Lisbon, Portugal.

B. Dorr, D. Zajic, and R. Schwartz. 2003. Hedge Trimmer: a parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL Text Summarization Workshop and Document Understanding Conference*, pages 1–8, Edmonton, Canada.

N. Hatzigeorgiu, M. Gavrilidou, S. Piperidis, G. Carayannis, A. Papakostopoulou, A. Spiliotopoulou, A. Vacalopoulou, P. Labropoulou, E. Mantzari, H. Papageorgiou, and I. Demiros. 2000. Design and implementation of the online ILSP Greek Corpus. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 1737–1742, Athens, Greece.

A. Iordanidou. 2006. *Thisavros Sinonimon ke Antitheton tis Neas Elinikis*. Patakis, Athens, Greece.

F. Karamitroglou. 1998. A proposed set of subtitling standards in Europe. *Translation Journal*, 2(2). http://accurapid.com/journal/04stndrd.htm.

K Knight and D. Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *Proceedings of the 17th National Conference of the American Association for Artificial Intelligence*, pages 703–710, Austin, USA.

J. Nivre, J. Hall, and J. Nilsson. 2004. Memory-Based Dependency Parsing. In *Proceedings of the 8th Conference on Computational Natural Language Learning*, pages 49–56, Boston, USA.

H. Papageorgiou, P. Prokopidis, I. Demiros, V. Giouli, A. Konstantinidis, and S. Piperidis. 2002. Multi-level XML-based Corpus Annotation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

P. Prokopidis, E. Desypri, M. Koutsombogera, H. Papageorgiou, and P. Piperidis. 2005. Theoretical and Practical Issues in the Construction of a Greek Dependency Treebank. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories*, pages 149–160, Barcelona, Spain.

S. Riezler, T.H. King, R. Crouch, and A. Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for Lexical-Functional Grammar. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 118–125, Edmonton, Canada.