Summarizing Dutch Sentences for Automatic Subtitling

Erik F. Tjong Kim Sang
CNTS – Language Technology Group
University of Antwerp – Belgium
erik.tjongkimsang@ua.ac.be

Project Summary

- ATraNoS: Automatic Transcription and Normalization of Speech
- Sponsor: IWT
- Goal: generate better products for automatically transcribing speech
- Case study: generating teletext subtitles for hearing-impaired people
- Research partners: Ghent University, Katholieke Universiteit Leuven, University of Antwerp

Sentence summarization

In the subtitle space on the tv screen often there is not enough room for displaying the sentences generated by the speech recognizer.

If a sentence does not fit, it needs to be summarized.

We summarize sentences in a two-step process:

1. make a shallow syntactic analysis of the sentence
2. determine what parts of the sentence can be removed or replaced

Both tasks are performed by machine learning algorithms (memory-based learners) trained on examples.

Shallow syntactic analysis modules: examples (1)

1. Part-of-speech tagger
   basketballer/N(soort,ev,basis,zijd,stan) Dennis/SPEC(deeleigen) Rodman/SPEC(deeleigen) heeft/WW(pv,tgw,met-t) van/VZ(init) zich/VNW(refl,pron,obl,red,3,getal) laten/WW(inf,vrij,zonder) horen/WW(inf,vrij,zonder) ./LET()

2. Lemmatizer
   basketballer Dennis Rodman heeft/hebben van zich laten horen .

Shallow syntactic analysis modules: examples (2)

3. Text chunker
   [NP basketballer NP] [MWU Dennis Rodman MWU]
   [VP-1 heeft VP-1] [PP van PP] [NP zich NP]
   [VP-1 laten horen VP-1] .

4. Relation finder
   [SU basketballer Dennis Rodman SU] [HD heeft HD] van zich laten horen .

5. Person name finder
   basketballer [FIRST Dennis FIRST] [SUR Rodman SUR] heeft van zich laten horen .

Shallow syntactic analysis modules: data

Nearly all modules use material from the Spoken Dutch Corpus (CGN).

Tagging and lemmatization were performed with 5,863,440 tokens of training data and 59,869 tokens of test data.

Chunking and relation finding used 369,023 tokens of training data and 45,533 tokens of test data.

Training and test data for the person name experiments was obtained from the CoNLL-2002 named-entity recognition task (182,796 tokens of training material and 20,135 tokens for testing).

Shallow syntactic analysis modules: performances

task	baseline	TnT	MBL
POS tagging	90.1	96.9	96.5
Lemmatization	98.2	98.9	99.0
Chunking	84.4	90.8	92.8
Relation finding	88.7	88.5	92.7
Name finding	74.8	63.0	86.0

The table contains $F_{\beta=1}$ rates.

We work on additional modules for processing speech recognizer output: punctuation insertion (MBL: 40.2) and capitalization.

Summarization: approach

We have defined the summarization task as a word classification task. Words can be copied, deleted or replaced:

politici vinden de euro natuurlijk/DELETE een goeie/goede zaak

The performance of the system was measured with precision, recall and $F_{\beta=1}$ rates for deletions and replacements according to the corpus.

Additionally the compression rate was measured (percentage of characters left after summarization).

Summarization: corpora

We have three parallel corpora of transcripts and subtitles:

corpus	**size** (words)	**compression rate** (chars)
VRT (news)	431,190	78%
NOS (news)	213,176	73%
Thuis (soap)	20,387	86%
Used (VRT)	92,875	81%

Many sentence pairs in the parallel corpus consist of copies. These pairs are not very interesting as summarization examples. Therefore only a part of the corpus has been used in the experiments.

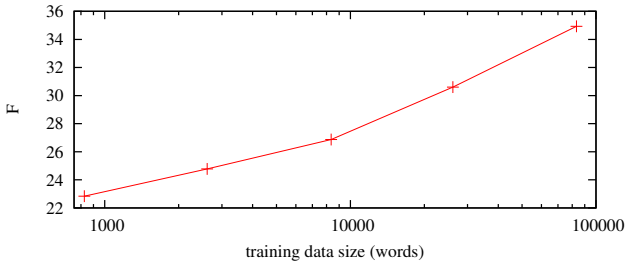
Summarization: results (1)

method	data type	comp.r.	precision	recall	$F_{\beta=1}$
baseline	words	96%	53.3%	13.6%	21.7
MBL	all	86%	41.2%	30.3%	34.9

The target compression rate was 81%. Performance was measured by counting the number of correct word replacements and word deletions.

After feature selection, the memory-based learner employs words, lemmas, part-of-speech tags and person name information, but neither chunk information nor relation information.

Relation training data size - performance



This learning curve suggests that the performance would improve significantly if we would have access to more training data.

Alternative summarization approach

We have evaluated an alternative approach which employs hand-crafted word and phrase deletion rules.

These rules select words and phrases which can be deleted from a sentence.

They have access to words, lemmas, part-of-speech tags, chunk tags and person name information.

When the selection process is complete, another process decides what words and phrases to delete, based on the required compression rate.

Summarization: results (2)

method	data type	comp.r.	precision	recall	$F_{\beta=1}$
baseline	words	96%	53.3%	13.6%	21.7
MBL	all	86%	41.2%	30.3%	34.9
Rules	all	76%	28.7%	28.9%	28.8

The target compression rate was 81%.

The rule-based approach performs worse than the learning approach with respect to simulating the choices of the human summarizer.

However, the summarized sentences produced by the rule-based approach meet the compression rate requirements more often.

Example output (memory-based learner)

Marc Dutroux wil onder de huidige voorwaarden niet naar $\{_1$ zijn(z'n) $\}$ proces . Opnieuw is een toren van de cokesfabriek Marly platgelegd . President Chirac wil geen hoofddoeken op de Franse scholen . De luchtvaart viert z'n honderdste verjaardag . [$_2$ en] Straks [$_3$ kennen we] de sportman en sportvrouw van het jaar . $\{_4$ het('t) $\}$ Journaal van zeven [$_5$ uur] met Martine Tanghe . Goeieavond . Marc Dutroux zegt dat hij niet naar z'n proces wil gaan als hij een bivakmuts moet opzetten tijdens $\{_6$ zijn(z'n) $\}$ transport van de [$_7$ gevangenis] naar de assissenzaal . $\{_8$ Zijn(Z'n) $\}$ advocaten vinden die [$_9$ bijkomende] veiligheidsmaatregel ongehoord [$_{10}$ en] ze zijn van plan [$_{11}$ om] over de kwestie een brief te schrijven [$_{12}$ naar] de minister [$_{13}$ van Binnenlandse Zaken] [$_{14}$ en die van] Justitie .

Example output (handcrafted deletion rules)

Marc Dutroux wil onder de huidige voorwaarden niet [$_1$ naar z'n proces] . Opnieuw is een toren [$_2$ van de cokesfabriek Marly] platgelegd . President Chirac wil geen hoofddoeken [$_3$ op de Franse scholen] . De luchtvaart viert z'n [$_4$ honderdste] verjaardag . En straks kennen we de sportman [$_5$ en sportvrouw] van het jaar . 't Journaal van [$_6$ zeven] uur met [$_7$ Martine] Tanghe . Goeieavond . Marc Dutroux zegt dat hij niet naar z'n proces wil gaan als hij een bivakmuts moet opzetten tijdens z'n transport [$_8$ van de gevangenis naar de assissenzaal] . Z'n advocaten vinden die bijkomende veiligheidsmaatregel ongehoord en ze zijn van plan om over de kwestie een brief te schrijven [$_9$ naar de minister van Binnenlandse Zaken] en die van Justitie .

Concluding remarks

We have examined two methods for summarizing Dutch sentences.

Both approaches benefited from having access to a shallow syntactic analysis of the input sentences.

The machine learning approach does not delete enough words to meet the required compression rates, probably due to a lack of training data.

Handcrafted deletion rules generate sentences that are short enough.

A more elaborate syntactic analysis is necessary to get rid of some of the summarization errors.

Demo

The most recent version of the two summarizers can be tested at

http://cnts.uia.ac.be/cgi-bin/atranos

This site can be found by searching for tjong and demo on Google.