Summarizing Sentences for Automatic Subtitling

Erik Tjong Kim Sang
CNTS – Language Technology Group
University of Antwerp

ATraNoS

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

Summarizing sentences

Problem: there is not enough room for displaying text generated by the speech recognition system.

Solution: summarize the text.

Method: memory-based: summarized sentences will be built from a corpus of examples.

Example

- Over deze man gaat het , Noël Slangen , communicatieadviseur van premier Verhofstadt , en eigenaar van een communicatiebedrijf .
- Noel Slangen is communicatie-adviseur van Verhofstadt .
- Hij heeft een communicatiebedrijf

Method

- Collect relevant data (teletext subtitles and associated transcribed programmes).
- Align data on sentence level and check.
- Annotate data with shallow syntactic structure.
- Generate phrase or word alignment (no check).
- Build sentence summarization system based on examples in corpus.

Corpus size

Corpus	Parallel	Compr.rate	Subtitles
VRT	189,983	76.6%	993,102
NOS	213,176	73.4%	-
Thuis	20,387	85.7%	200,358

Sizes have been measured in number of words (strings containing at least one of [A-Za-z0-9]).

Automatic alignment

- The standard method for aligning translated sentences (Gale and Curch) does not work well because often subtitle sentences are missing.
- We used a method which estimates the probability that sentences belong together based on the words they contain.
- The system benefits from making several passes over the data.

Alignment software performance

Corpus	Precision	Recall	$F_{ß=1}$
VRT	93.5%	85.3%	89.2
NOS	88.0%	87.2%	87.6
Thuis	81.5%	57.6%	67.5

Precision and recall have been computed for pairs of sentences. $F_{ß=1}$ is the harmonic mean of these.

Syntactic annotation

- A part-of-speech tagger was built by using the CGN release 5 material (4.1 Mwords).
- A baseNP chunker was built from the same corpus (0.3 Mwords).
- We used the memory-based tagger Mbt (available from http://ilk.uvt.nl).

Word alignment

- Similar words in at comparable positions in the same alignment stucture have been combined.
- Words with the same part-of-speech tag at comparable positions have been combined.
- Only 1→1, 1→0 and 0→1 alignments have been used.

Memory-based summarization

- A memory-based learner was trained to reproduce the subtitles from the broadcast transcriptions.
- For each transcribed word it performed one out of three actions: copy, delete or replace.
- Different features were evaluated: words, part-of-speech tags, predictions for context words.
- We used the Timbl software package (available from http://ilk.uvt.nl).

Results

Corpus	Baseline	Best	Features	Compr.r.
VRT	54.3%	60.4%	$w_0p_0p_{+1}$	71.6%
NOS	49.1%	57.9%	$w_0p_{-1}p_0$	53.1%
Thuis	52.1%	60.6%	$w_0p_0n_0n_{+1}$	66.4%

Features: w: words
p: part-of-speech tags
n: noun phrase tags.

Output example

Welkom bij het zeven uurjournaal . Dit zijn de hoofdpunten : In Afghanistan veroveren anti-Talibantroepen de tunnels van Tora Bora , maar van Osama bin Laden is geen spoor . Yasser Arafat roept de Palestijnse extremisten op de aanslagen tegen Israël te stoppen . En de Belgen grijpen naast de medailles op de Europese zwemkampioenschappen in Antwerpen . In Afghanistan hebben anti-Talibanstrijders de tunnels en grotten van Tora Bora veroverd , het laatste bolwerk van terroristenleider Osama bin Laden . Vanmiddag kwam één van de commandanten van de Alliantie terug van het front , met goed nieuws en slecht nieuws . Het goede was : Tora Bora is gevallen en Al Qaeda is verslagen .

Concluding remarks

- A parallel, sentence-aligned corpus of broadcast transcriptions and teletext subtitles has been built.
- A basic sentence summarization system was created based.
- Some of the predictions made by the system are correct.
- Future work: generate a more elaborate syntactic annotation for the corpus.