Compounds, Coreferences and Multiword Translations

Martin Volk, Laura Mascarell and Mark Fishel

Sinergia Project

MODERN: Modeling discourse entities and relations for coherent machine translation (2013 - 2016)
Partners: Idiap Research Center, Martigny (CH), University of Geneva (CH), University of Zurich (CH), University of Utrecht (NL)

Problem

Current MT systems translate each sentence in isolation. Many ambiguities can only be resolved by intersentential context.

Example

The English pronoun *it* must be translated into German as either *er, sie* or *es* depending on the grammatical gender of the antecedent.

Example

The English noun *face* must be translated into German as either *Wand* or *Gesicht* depending on the context. If used in a compound like *East face*, the translation will be *Wand*. [3] [2]

Sentence 1: *: . . . on the unclimbed* **East face** *of the Central Tower ...*

Sentence 2: *. . . we were swept from the* **face** *by a five-day storm . . .*

Our Corpus

We use the **Text+Berg corpus** with 5 million words of Alpine texts in German and French taken from yearbooks of the Swiss Alpine Club (1957-2013).[1]:

Example from Die ALPEN, Les ALPES, 1979:

German	French
Am 6. September versuchten Ridgeway und Roskelley den Gipfel direkt über die **Nordostwand** *anzugehen ...*	*Le 6 septembre 1978, Ridgeway et Roskelley tentèrent d'achever l'ascension directe de la* **face NE**, *...*
... überzeugt, dass die **Wand** *bei diesen Bedingungen eine sehr grosse Lawinengefahr barg.*	*... convaincus que les conditions de la* **face** *présentaient un sérieux danger d'avalanche.*

Solution

Statistical Machine Translation with caching for German to French MT. Solution via coreference detection and translation caching

1. Split German compound noun into segments $s_1, s_2, ..., s_n$ (e.g. *Nordostwand* $\rightarrow$ *Nord+ost+wand*). Save the translation of the last segment in a cache (e.g. DE: *Wand* – FR: *face*).

2. Use the last segment for co-reference detection in 4 subsequent sentences. If found, use the previous translation of the segment.

Source	Die Originalauswertung wurde in den Zwischenmassstab 1:20000 reduziert, worauf das ***Bundesamt*** (trans: *office fédéral*) für Landestopographie in Aktion trat. Nur dieses ***Amt*** war in der Lage, [...]
English translation by the authors	"The original evaluation was reduced in the intermediate scale 1:20000, followed by the *Federal Office* of Topography went into action. Only this *office* was able to [...]"
SMT-1, SMT-split	que ce ***poste*** était dans la situation, [...]
SMT-1 enf., SMT-split enf.	que de cet ***office*** était en mesure [...]

Example where our enforcing method improves the translation of the noun coreference.

Results

We assess the performance of our method in both approaches (i.e. splitting compounds and not splitting them), so we build two phrase-based SMT systems *SMT-1* and *SMT-split*. Both systems are built using the standard settings, 5-gram language model KenLM, and GIZA++. The language model is trained on a total of 624,160 sentences (13 million tokens), and the training set consists of 219,187 sentences and roughly 4.1/4.7 million words in German and French, respectively.

	Consistent: yes	Consistent: no
Correct	52	117
Incorrect	6	36

Consistency and correctness results of the ***SMT-1*** system without enforcing consistency.

	Consistent: yes	Consistent: no
Correct	73	102
Incorrect	7	29

Consistency and correctness results of the ***SMT-1*** system when our translation enforcing method is applied.

	Consistent: yes	Consistent: no
Correct	103	80
Incorrect	6	22

Consistency and correctness results of the ***SMT-split*** system when our translation enforcing method is applied.

	Correctness	Consistency
SMT-1	80.1%	27.5%
SMT-split	82.0%	35.1%
SMT-1 enf.	82.9%	37.9%
SMT-split enf.	86.7%	52.1%

Overall percentages of consistency and correctness results of the *SMT-1* and *SMT-split* systems, with and without our enforcing method.

Since the splitting method allows the *SMT-split* system to translate out-of-vocabulary compounds, *SMT-split* also increases the number of the enforced examples, improving the translation in a number of cases.

References

[1] Noah Bubenhofer, Martin Volk, Fabienne Leuenberger, Manuela Weibel, and Daniel Wüest, editors. *Text+Berg-Korpus (Release 149 v01). Digitale Edition des Jahrbuch des SAC 1864-1923, Echo des Alpes 1872-1924 und Die Alpen, Les Alpes, Le Alpi 1925-2013.* Institut für Computerlinguistik, Universität Zürich, 2014.

[2] Marine Carpuat. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27, Boulder, Colorado, 2009.

[3] Laura Mascarell, Mark Fishel, Natalia Korchagina, and Martin Volk. Enforcing consistent translation of German compound coreferences. In *Proceedings of KONVENS*, Hildesheim, 2014.

PARSEME, Malta, 19-20 March 2015
This research was supported by the Swiss National Science Foundation under grant CRSII2_147653/1 through the project "MODERN: Modelling discourse entities and relations for coherent machine translation".

Contact

University of Zurich[UZH]

Institute of Computational Linguistics
http://www.cl.uzh.ch

Martin Volk
Binzmühlestrasse 14, CH-8050 Zurich
volk@cl.uzh.ch