Linguistics, German Compounds and Statistical Machine Translation. Can they all get along?

Carla Parra Escartín, Stephan Peitz, Hermann Ney
University of Bergen, RWTH Aachen University
PARSEME WG3

Object of study

- Compounds are a known challenge both in Machine Translation (MT) and Translation in general as well as in other Natural Language Processing (NLP) applications.
- Impact of using linguistics to preprocess German compounds prior to translation in Statistical Machine Translation (SMT).

German Compounds

(1) *Warm* *Wasser* *Bereitung*
caliente agua preparación
warm water production
[ES]: 'Preparación de agua caliente'
[EN]: 'Warm water production'

(2) *Wärme* *Rückgewinnung* *s* *Systeme*
calor recuperación Ø sistemas
heat recovery Ø Systems
[ES]: 'sistemas de recuperación de calor'
[EN]: 'heat recovery systems'

- warm (ADJ) + Wasser(N) = **Warmwasser** (N) + Bereitung(N) = **Warmwasserbereitung** (N) + s + Anlagen(N) = **Warmwasserbereitungsanlagen** (N)
  [*EN: warm water production systems*]
- abstellen(V) - en + Anlagen(N) = Abstellanlagen (N)
  [*EN: parking facilities*]

Corpus statistics

	training	dev	test
Sentences	1.8M	2382	1192
Tokens	40.8M	20K	11K
Types	338K	4050	2087

- The training corpus is a concatenation of an internally compiled version of the Europarl Corpus [1] German→Spanish and a greater part of the TRIS corpus.
- dev and test only include texts from the TRIS corpus.

Experiments

As SMT system, we employ the state-of-the-art phrase-based translation approach [6], implemented in *Jane* [5].

1. *Baseline* 1: normalized version of a concatenation of the TRIS corpus and Europarl
2. *compList*: corpus + compound list
3. *RWTH* (baseline 2): compounds split
4. *RWTH + compList*: compounds split + compound list
5. *IMS*: compounds split
6. *IMS + compList*: compounds split + compound list

Conclusion

- The concatenation of compound lists to the training corpora seems to have had a positive impact in the overall results.
- Automatic ways of retrieving such lists would be desired.
- The best approach seems to be a combination of splitting the compounds and adding compound lists to training.

Acknowledgements

The current research was financed by the EU under the Marie Curie Actions, FP7 People programme (grant agreement 238405).

Case Study

	Text A	Text B
Number of words	2431	439
Number of comp.	265 (10.9%)	62 (14.12%)
Number of unique comp.	143	25
Lexicalized comp.	99 (4.07%)	18 (4.1%)
Unique lexicalized comp.	63	4
Not lexicalized comp.	166 (6.8%)	44 (10.06%)
Unique not lexicalized comp.	80	21

Compound nominals found in two texts taken from the TRIS corpus [2].

Compounds detected

	Popović et al. (2006)	Weller and Heid (2012)
Compounds in training	182334	141789
% Vocabulary	54%	42%
% Running words	0.4%	0.3%
Compounds in test	924	444
% Vocabulary	44.3%	21.3%
% Running words	8.5%	4%

Compounds detected by each of the splitters and percentages they account for with respect to the vocabulary and the number of running words in the corpora used in the experiments.

Results for the German→Spanish TRIS data

Experiment	Splitting Method	test BLEU [%]	test TER [%]	test OOVs
Baseline	-	45.9	43.9	181
compList	-	**46.7**	**42.9**	169
RWTH	Popović et al. (2006)	48.3	40.8	104
RWTH+compList		**49.1**	40.5	104
IMS	Weller and Heid (2012)	48.3	40.5	114
IMS+compList		**49.7**	**39.2**	114

Future work

1. Replicate the experiments with data used by the MT community.
2. Replicate the experiments for other language pairs.

→ Obtaining positive results in these further experiments would suggest that a similar approach may also yield positive results in dealing with other types of MWEs within SMT.

References

[1] Philipp Koehn, *Europarl: A Parallel Corpus for Statistical Machine Translation*, Conference Proceedings: the Tenth Machine Translation Summit (Phuket, Thailand), 2005, pp. 79–86.

[2] Carla Parra Escartín, *Design and compilation of a specialized Spanish-German parallel corpus*, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (Istanbul, Turkey), European Language Resources Association (ELRA), May 2012, pp. 2199–2206 (English).

[3] Maja Popović, Daniel Stein, and Hermann Ney, *Statistical machine translation of german compound words*, Proceedings of the 5th international conference on Advances in Natural Language Processing (Berlin, Heidelberg), FinTAL'06, Springer-Verlag, 2006, pp. 616–624.

[4] Marion Weller and Ulrich Heid, *Analyzing and Aligning German compound nouns*, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (Istanbul, Turkey), European Language Resources Association, May 2012.

[5] Joern Wuebker, Matthias Huck, Stephan Peitz, Malte Nuhn, Markus Freitag, Jan-Thorsten Peter, Saab Mansour, and Hermann Ney, *Jane 2: Open source phrase-based and hierarchical statistical machine translation*, International Conference on Computational Linguistics (Mumbai, India), December 2012, pp. 483–491.

[6] Richard Zens and Hermann Ney, *Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation*, International Workshop on Spoken Language Translation (Honolulu, Hawaii), October 2008, pp. 195–205.