Approximating compound compositionality based on word alignments

Fabienne Cap
Uppsala University
fabienne.cap@lingfil.uu.se

Introduction We approximate the compositionality of German noun-noun compounds using statistical word alignments, based on (Villada Moirón and Tiedemann, 2016). Our hypothesis is that compositional constructions are translated similarly by human translators, whereas non-compositional constructions exhibit more variance. When training a statistical word alignment this greater variance leads to a large number of different alignments, which we use to determine the compositionality of a construction.

Experimental Setup We split all noun-noun compounds occuring in the German Europarl corpus (Koehn, 2005) and then run statistical word alignment on the English and the modified German corpus. We then calculate the *translational entropy* (TE) score (Villada Moirón and Tiedemann, 2016) and sort the compounds in descending order so that compounds with the greatest likelihood of being non-compositional appear at the top of the list. First, the TE-scores of both components are weighted equally, but different weightings are investigated. More lists are produced, sorted according to the TE-score of either modifiers or heads.

Results In Figure 1(a) we show some examples from our lists with the modifier *Auge*, which show that TE scores correlate well with compositionality. 1(b) illustrates the greater variance in the non-compositional *Augenzwinkern* compared to *Augenschäden*. Figure 1:

Compound	TE
Auge|Maß	3.428
Auge|Höhe	2.236
Auge|Zwinkern	1.748
Auge|Schäden	0.637

(a) TE scores.

Word		Alignments
Auge (Zwinkern)	=	nod (2), cheek (1), a (1), glint (1), blind eye (1), personalise (1)
Auge (Schäden)	=	eye (3)

(b) Word alignments for *Auge*.

References: • Koehn, P. (2005): Europarl: a parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*. • Villada Moirón, B. and Tiedemann, J. (2006): Identifying idiomatic Expressions using automatic word alignment. In: *Proceedings of the EACL 2006 MWE Workshop*.