Using Word Alignments to Determine the Compositionality of Swedish Compound Nouns

Fabienne Cap and Sara Stymne

Department of Linguistics and Philology
Uppsala University

1. Introduction

We present an approach to approximate the compositionality of Swedish noun-noun compounds using statistical word alignments. It is based on previous work by Villada Moirón and Tiedemann (2006), who used word alignments to identify non-compositional multiword expressions in Dutch. The underlying 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 (eventually erroneous) alignments, which in turn can be identified and used to determine the compositionality of a construction. An example is given in Figure 1 In the seven occurrences of the semi-compositional word *grunddrag*, the modifier *grund* is aligned to six different words and the head *drag* to four different words. In contrast, in the seven occurrences of the fully compositional *grundforskning*, the modifier *grund* is aligned to only three different words and the head *forskning* only to two.

Word		**Alignments**
grund	=	basic (2), the key (1), feature (1), the (1), fundamental (1), main (1)
drag	=	features (3), outline (1), feature (1), lines (1)

Word		**Alignments**
grund	=	basic (5), engineering (1), the (1)
forskning	=	research (6), engineering (1)

Figure 1: Alignments for the semi-compositional word *grunddrag* (TE: 1.748) and the fully compositional word *grundforskning* (TE: 0.603).

These differences in alignment variance can be expressed by using a *translational entropy* score (Villada Moirón and Tiedemann, 2006). In the following, we will report on how we adapted their approach to identify non-compositional Swedish noun-noun compounds.

2. Experimental Setup

We base our experiments on the Swedish Europarl corpus (Koehn, 2005), with a total of 1,825,809 sentences. We split all occurring Swedish noun compounds using a combined corpus- and POS-based method as described in (Stymne and Holmqvist, 2008), which is an extension of the method proposed by Koehn and Knight (2003). We try all possible splits for each noun, allowing a few morphological changes like the addition of an *s*, and choose the split with the highest arithmetic mean of the frequencies of its parts. We only allow modifiers and heads that have occurred as nouns in the corpus. For tagging we used Granska (Carlberger and Kann, 1999). For our compositionality experiments we only consider compounds that are split into two parts, meaning that we ignore splits like *järn+vägs+nät*, but allow *andrabehandlings+rekommendation* even though the first part could have been split into two parts. For our analysis, we ignore compounds that occur less than 10 times in the corpus.

After splitting the corpus, we run statistical word alignment (Gao and Vogel, 2008) on the English and the modified Swedish section of Europarl, and calculate the *translational entropy* (TE) scores as described in (Villada Moirón and Tiedemann, 2006), and shown in equation 1, where T_s is the compound with its two parts, $P(t|s)$ is the proportion of alignment t among all alignments of the word s in the context of the given compound.

$$H(T_s|s) = -\sum_{t \in T_s} P(t|s) \log P(t|s) \qquad (1)$$

We then rank the compounds in descending order of this score so that compounds with the greatest likelihood of being non-compositional appear at the top of the list while compositional compounds occur at its bottom.

3. Results

We compare our ranking to two baselines: one in which the compounds are ranked according to their frequency and one in which they are ranked according to the mean frequency of their parts. We then annotated the top 25 of each of these rankings with respect to their compositionality into three groups: i) compositional compounds, ii) semi-compositional compounds and iii) non-compositional compounds. The result is given in Figure 2. It can be seen that the TE-based ranking yields more non-compositional (e.g. *ståndpunkt*) and semi-compositional (e.g. *tidpunkt*) compounds at the top of the list than the two baselines. This is an indicator that the method is applicable not only to determine idiomatic multiword expressions (Villada Moirón and Tiedemann, 2006), but also closed noun-noun compounds.

Moreover, the list also reveals errors of the compound splitter, e.g. *handel*, which the splitter considered as a split of *hane* and *del*.

Following Villada Moirón and Tiedemann (2006), we compare the top list as ranked by TE to excerpts from the middle and bottom of the list. In both these samples, shown in Figure 3, we find a few semi-compositional compounds, but no non-compositional compounds. This supports our

Compounds ranked by frequency	Freq	Compounds ranked by part frequency	PF	Compounds ranked by TE	TE
ändringsförslag	18,142	landkommission	127250	ståndpunkt	5.973
kommisionsledamot	11,827	landfråga	118659	*handel	5.836
ståndpunkt	11,308	kommissionsförslag	116960	synsätt	5.668
jordbrukspolitik	5,039	medlemsstatsfråga	114034	synpunkt	5.572
*handel	5,019	parlamentsfråga	113898	*fördel	5.556
*fördel	4,242	*rådfråga	112655	ändringsförslag	5.494
*framgång	4,182	unionsfråga	112599	tidpunkt	5.434
synpunkt	3,959	kommissionsåtgärd	107713	ställningstagande	5.419
arbetstillfälle	3,880	kommissionsarbete	98937	parlamentsledamot	5.319
parlamentsledamot	3,763	rättighetsfråga	98870	*målsättning	5.310
konkurrenskraft	3,575	*herrtalman	98698	livsmedel	5.157
handlingsplan	3,230	kommissionsordförande	98682	näringsliv	5.104
rådsordförande	3,158	kommissionsdirektiv	95372	förhållningssätt	5.081
*målsättning	3,123	kommissionsbeslut	94289	*deltagande	5.041
medlemsland	3,094	kommissionspolitik	93946	tjänsteman	5.008
folkhälsa	2,952	kommissionsledamot	93682	nätverk	4.941
deltagande	2,833	regeringskommission	93480	kommisionsledamot	4.918
resolutionsförslag	2,748	stödfråga	91721	utgångspunkt	4.895
*kommissionären	2,714	världskommission	91064	*föremål	4.877
rättvisa	2,659	kommissionsordförandeskap	90298	ordalag	4.874
folkparti	2,561	utvecklingsfråga	89966	underlag	4.865
tidpunkt	2,411	debattfråga	89924	ändamål	4.855
regelverk	2,213	unionland	89847	*framgång	4.826
säkerhetspolitik	2,162	omröstningskommission	89711	tyngdpunkt	4.765
arbetsmarknad	2,128	kommissionslagstiftning	89046	regelverk	4.745

Figure 2: Top 25 noun-noun compounds, sorted in descending frequency, part frequency and translational entropy score. Erroneous splits are marked with an asterix (*). Compositional compounds do not bear markup, e.g. *ändringsförslag*. Semi-compositional compounds are marked grey e.g. *regelverk*. Non-compositional compounds are highlighted darker, e.g.*synpunkt*.

hypothesis that TE scores are useful to determine the compositionality of noun-noun compounds.

In addition to excerpts of the full list given in Figures 2+3, we also extracted a sublist that shares the same modifier in Table 1. This list shows nicely how the compositionality of the compounds including *hand* increases as the TE score decreases.

Compound	TE
hand|läggning	3.925
hand|ledning	2.607
hand|bok	2.139
hand|bagage	1.773
hand|tag	1.761
hand|väska	1.461
hand|verktyg	1.386

Table 1: TE scores for compounds with *hand*.

4. Conclusion and Future Work

In conclusion, based on the evidences from Figures 2+3 and Table 1 we can say that the approach of Villada Moirón and Tiedemann (2006) which has been successfully applied to multiword expressions in the past is also working to rank closed noun compounds according to their compositionality. In the future, we plan to perform a more detailed analysis of the results. Moreover, we want to extend the alignment approach to align the Swedish section not only to English but also to other languages in order to obtain scores that are independent of eventual similarities or other peculiarities between the two languages used.

Middle Excerpt		Bottom Excerpt	
Compound	TE	Compound	TE
persondator	2.398	presstjänst	0.760
nationstat	2.397	vägtrafikanter	0.752
sysselsättningsriktlinje	2.397	kustbevakningsenhet	0.750
valkommission	2.397	uppvärmningspotential	0.744
omstruktureringsstöd	2.397	kopplingsdirektiv	0.743
kontrollbesök	2.396	bolagsstadga	0.737
giltighetsområde	2.396	sockerpolitik	0.733
folkstyre	2.396	momsstrategi	0.724
forskningsområde	2.396	kärnfusion	0.724
betalningsvillkor	2.395	utvidgningsvåg	0.721
kvalitetsmärkning	2.394	fredskår	0.718
statsförvaltning	2.394	utlåningskapacitet	0.717
klimatfråga	2.394	sälprodukt	0.704
gemenskapsbidrag	2.394	konvergensrapport	0.701
förbundsstat	2.394	systerparti	0.680
lägenhet	2.393	könsidentitet	0.639
förhandlingsposition	2.393	utvecklingsprioritering	0.632
gemenskapsmodell	2.392	sharialagstiftning	0.612
anslutningsstrategi	2.392	privatföretag	0.600
exportprodukt	2.391	skiffergas	0.561
kreditgivning	2.391	partistadga	0.555
säkerhetsstandard	2.391	kalenderår	0.516
tjänsteavdelning	2.391	industriprodukt	0.509
kärnenergiprogram	2.391	interventionsplan	0.482
utskottsledamot	2.391	röstavsikt	0.104

Figure 3: Middle and bottom ranks of the TE-list.

References

Johan Carlberger and Viggo Kann. 1999. Implementing an efficient part-of-speech tagger. *Software Practice and Experience*, 29:815–832.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *ACL'08: Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 49–57. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the EACL*, pages 187–193, Budapest, Hungary.

Philipp Koehn. 2005. Europarl: a Parallel Corpus for Statistical Machine Translation. In *MT Summit'05: Proceedings of the 10th machine translation summit*, pages 79–86.

Sara Stymne and Maria Holmqvist. 2008. Processing of Swedish Compounds for Phrase-based Statistical Machine Translation. In *EAMT '08: Proceedings of the 12th annual conference of the European Association for machine translation*, pages 180–189.

Begońa Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on multi-word-expressions in a multilingual context*, pages 33–40.