Automatic Translation of Norwegian Noun Compounds

Lars Bungum
Department of Informatics
University of Oslo
larsbun@ifi.uio.no

Stephan Oepen
Department of Informatics
University of Oslo
oe@ifi.uio.no

Abstract

This paper discusses the automated translation of Norwegian nominal compounds into English, combining (a) compound segmentation, (b) component translation, (c) bi-lingual translation templates, and (d) probabilistic ranking. In this approach, a Norwegian compound will typically give rise to a large number of possible translations, and the selection of the 'right' candidate is approaches as an interesting machine learning problem. Our work extends the seminal approach of Tanaka and Baldwin in several ways, including a clarification of some fine points of their earlier work, adaptation to a more adequate machine learning framework, application to a Germanic language with a small speech community and very limited existing resources, and systematic experimentation along several dimensions of variation.

1 Background: The Task

Compounding is a productive feature of the Norwegian language (just as in other Germanic language), and because Norwegian compounds are written in a single word (i.e. as one blank-separated entity) such constructions pose a challenge to automatic translation.[1] Consider the examples in (1), where we use a centered dot ('·') to typographically indicate component boundaries both in Norwegian compounds and literal English glosses:

(1) a. anlegg·*s*·vei
construction·road
'construction road'

b. dokument·stabel
document·pile
'pile of documents'

c. brud·*e*·spore
bride·spur
'fragrant orchid'

Both examples (1-a) and (1-b) can be translated adequately from the translations of their component parts: in (1-a) the formative *-s-* joins together the two components, where in (1-b) the Norwegian compound merely is the juxtaposition of two independent 'words'.[2] In terms of aligning components during translation, the Norwegian surface order is preserved in (1-a) (the English translation being a regular noun – noun compound), while (1-b) reverses the order of the component parts—in a different English construction, using the prepositional marker *of*.[3] We will refer to the correspondences between compound parts across languages as *translation templates* (see Section 3 below), where (1-a) and (1-b), for example, instantiate the templates $\langle N_1 N_2 \rangle \rightarrow \langle E_1 E_2 \rangle$ and $\langle N_1 N_2 \rangle \rightarrow \langle E_2 \text{ of } E_1 \rangle$, respectively.

Examples (1-a) and (1-b) are within the scope of our method, while (1-c) is not. The translation *fragrant orchid* is not accessible merely by translating the component parts of the Norwegian *brudespore*, and we call (1-c) *non-compositional* for our purposes. Furthermore, we limit our discussion to Norwegian *nominal* compounds with exactly *two* components, i.e. source language (SL) forms of the type $\langle N_1 N_2 \rangle$. We approach the task of translating such compounds as a processing pipeline of (a) compound analysis, (b) component translation, (c) template instantiation, and (d) ranking of translation candidates.

The number of candidate translations grows with the fertility of each component and the overall number of translation templates. We treat the selection of the best candidate as a ranking problem, employing a Maximum Entropy (MaxEnt) machine learning approach, and using a wide

[1] The Google translation services, for example, arguably present the best-performing open-domain Norwegian – English MT system to date. Nevertheless, the Google SMT system has no provisions for productively formed compounds.

[2] We use the term *word* in a purely technical sense here, i.e. for an independent unit of translation. In terms of the morphological structure of Norwegian compounds, the predominant analysis is as the combination of two (uninflected) *stems* (or lexemes), with inflection applying after compounding.

[3] For this example, it would seem appropriate to analyze *pile* as a relational noun, which would make the *of* PP a complement to the head noun. But for the purpose of the present discussion, nothing much will hinge on the specifics of the internal syntactic structure of English translations.

Proceedings of the 13th Annual Conference of the EAMT, pages 136–143,
Barcelona, May 2009

range of so-called *features*, encoding both mono-lingual and bi-lingual information for each translation candidate. Various MaxEnt ranking models are trained on a hand-crafted gold standard of 750 Norwegian compounds and preferred translations, and evaluated by means of cross-validation. Using this method, the best-performing model was able to select the exact gold standard translation for unseen test data in well above 50% of all cases.

In the following, we review closely related earlier work (Section 2), sketch the selection of experimental data, available resources, and specifics of our approach (Section 3), lay out the design of our experiments (Section 4), present a wealth of empirical results (Section 5), and finally conclude with a critical discussion of our findings (Section 6).

2 Earlier Work

In investigating the automatic translation of Norwegian nominal compounds, our starting point is the influential approach of Tanaka and Baldwin—henceforth T&B—who explore various ways of translating Japanese nominal compounds into English and vice versa (Tanaka and Baldwin, 2003a; Tanaka and Baldwin, 2003b; Baldwin and Tanaka, 2004). Abstractly, our steps (a) to (d) as sketched above are all taken from T&B, but there are important differences in the specifics of our approach, as well as extensions beyond the results of T&B. Besides, our focus on another language pair (with severely more limited resources available on the Norwegian source language side), most of the relevant differences pertain to the ranking step, arguably the key component in obtaining high-quality translations.

Tanaka and Baldwin (2003b) suggest to rank candidate translations based on target language (TL) distributional properties, essentially corpus frequencies. They develop an interpolated measure CTQ ('Corpus-based Translation Quality'; see Section 4 below), essentially ranking candidate translations according to the probabilities of component parts—relative to construction type, i.e. the English side of each translation template—and the probability of the candidate as a whole. CTQ is the reflection of linguistic arguments pointing to the importance both of the quantitative occurrence of a compound *itself* in a corpus, as well as to the propensity of its component parts to form phrases (of a specific construction type).

To avoid stipulating CTQ interpolation weights, Baldwin and Tanaka (2004) turn to a machine learning approach, proposing the creative (but mathematically dubious) use of a Support Vector Machine (SVM) classifier for the ranking task. The main shortcoming of their use of the SVM is the non-conditional nature of the probabilistic model, i.e. much like with CTQ; the task is construed as separating 'good' from 'bad' translations *independent* of the original SL compound.[4]

At the same time, Baldwin and Tanaka (2004) introduce additional sources of information, viz. bi-lingual properties extracted from machine-readable dictionaries. Intuitively, these additional machine learning features aim to provide a measure of the strength of the translation relation holding between component parts, and of course to actually capture those cases where SL compounds are fully listed in the dictionary. Our work extends Baldwin and Tanaka (2004) in several ways. First, we deploy a *conditional* MaxEnt ranker (rather than a contorted SVM classifier), leading to a formally more adequate and more scalable machine learning framework. We explore additional feature combinations of mono-lingual and bi-lingual sources of information, and provide a systematic investigation into the relevance of analysis 'depth' (contrasting a tagger vs. a syntactic parser) in preprocessing the training corpus. Finally, we provide empirical results on the learning curves—with increasing amounts of mono-lingual training data—of our various methods.

While T&B have been the foremost source of inspiration for our work, earlier approaches to the compound analysis and translation problem include Rackow et al. (1992), who explore the translation of German compounds into English. While their task is quite similar, this work has its emphasis on the segmentation and analysis of SL compounds, although it proposes using corpus data (counts) to distinguish between the various candidate translations. From the available information, the approach was not fully implemented or evaluated empirically. Grefenstette (1999), translating German and Spanish compounds, shows how WWW counts can be used to rank candidates, although his experiments are confined only to compounds for which a translation exists in a bi-lingual dictionary.

[4] Baldwin and Tanaka (2004) report that, in their SVM experiments, most of their training runs failed to converge, i.e. did not result in a functional classifier. This observation may well be owed to their creative use of the SVM framework.

3 Methodology and Preparational Steps

We pursued a data-driven approach both in the selection of training and test compounds and in the discovery of bi-lingual translation templates. A balanced set of 750 Norwegian $\langle N_1 N_2 \rangle$ compounds were extracted from running text, hand-inspected, and manually translated into English. Translation templates were then 'read off' the translations (the gold standard).

3.1 Source Language Compound Selection

Candidate Norwegian nominal compounds were selected from a large collection of running text, comprised of the Norwegian segments of the Oslo Multilingual Corpus[5], and of the smaller LOGON corpus (Oepen et al., 2004). The text corpus was analyzed using the Oslo-Bergen Tagger (OBT) (Hagen et al., 2000), which assigns a special SAMSET ('compound') tag to candidate compounds (i.e. input tokens not in the system lexicon, where a segmentation into known components is possible). Out of a total of 2,7 million words, 37,058 instances were labelled as compounds and nominals, of which 22,339 were unique types.[6]

To gauge frequency of use, Internet searches (using the Yahoo API) were performed for each of the unique compounds, and from the 4946 types that acquired more than 10 hits, we selected 750 at random. Much like in the original T&B experiments, these randomly chosen compounds were organized according to three frequency bands (according to Yahoo hits), henceforth: the low, middle and high bands. To identify compound-internal structure and confirm the $\langle N_1 N_2 \rangle$ construction type, we applied the procedure of Johannessen and Hauglin (1996), which is available as an optional component in the OBT. During this step, candidates that were segmented into more than two parts or other construction types were rejected and replaced with new random samples from the original set of 4946 words.

3.2 Gold Standard and Templates

Our final selection of 750 Norwegian $\langle N_1 N_2 \rangle$ compounds was presented to a bi-lingual informant, alongside the results of look-up in a Norwegian – English dictionary (Eek, 2001). The informant could either accept the translation, replace it or add to it, and provide translations for the compounds that were not listed in the available dictionary, which was the case for 95,6% of the compounds. Although alternatives in the translation were permitted, the informant was *not* instructed to provide an exhaustive list of possible translations. This was preferred to limiting the number of translations to one in all cases (as is the case in the earlier T&B experiments), as this would imply the undesirable assumption that any Norwegian compound, independent of context, has one and only one correct English translation. Of the 750 final SL compounds, 444 are compositional in our sense, i.e. the gold standard translation is available, in principle, to our method. The experiments reported in Section 5 focus on this compositional sub-set.

All translations were inspected and generalized into translation templates, essentially syntactic alignment instructions. The two templates seen earlier—$\langle N_1 N_2 \rangle \rightarrow \langle E_1 E_2 \rangle$ and $\langle N_1 N_2 \rangle \rightarrow \langle E_2 \text{ of } E_1 \rangle$—were the by far most frequent ones. We arrived at a total of 20 templates, including possessive constructions (e.g. *kvinne·avis* – woman·newspaper – 'woman's newspaper'), variation of the prepositional link (e.g. *jakt·lykke* – hunting·luck – 'luck in hunting'), morpho-syntactic variation of the non-head component, and even the reversed $\langle N_1 N_2 \rangle \rightarrow \langle E_2 E_1 \rangle$ (*gartner·mester* – gardener·master – 'master gardener'). This latter template which was attested only once in the gold standard, was excluded from our experiments as non-productive.

3.3 Target Language Statistics

A central element in the ranking of candidate translations is mono-lingual frequency information about the target language. To sample appropriate statistics, three large corpora of English text were used as the basis for the ranking task. The British National Corpus (BNC), comprising 80M words, the AQUAINT (AQ) corpus consisting of 375M words and finally the North American News Text Corpus (NAN) totalling 350M, words were all processed through the second version of the RASP parser (Briscoe et al., 2006), to make it possible to not only gather statistics of word (co-)occurrences but to also take into account the specific construc-

[5] See `http://www.hf.uio.no/ilos/OMC`.

[6] Note that these figures do not accurately reflect the frequency of compounding in Norwegian, as the OBT lexicon includes a relatively large number of high-frequency compounds, including many fully transparent and compositional ones. Due to the current OBT architecture, these instances are no longer identified with the SAMSET tag.

tion types. The parsed results were indexed according to the various templates, so that occurrence statistics for the compounds, their component parts, and the TL template structure could be easily extracted. In Section 4 below, we define various machine learning features on the basis of this data, and in Section 5, we investigate the effects of increasing amounts of available TL training data.

3.4 Task Definition and Evaluation

Our task is to automatically translate compounds according to the method outlined earlier. Seeing that the search space (the set of candidate translations) is fully determined by the bi-lingual dictionary and set of bi-lingual templates, the main factor of variation in our investigation is the ranking method applied to picking the 'best' candidate. In our experiments, we apply various rankers and evaluate against the gold standard translations. More precisely, we report the success rate as the percentage of Norwegian compounds for which the highest-ranked translation candidate is identical to the gold standard translation (or, in case of multiple references in the gold standard, is a member of that set). For the machine learning experiments, we apply ten-fold cross-validation, i.e. train the ranker on 90% of the gold standard and evaluate on the remaining 10%, repeating this procedure for all ten distinct splits, and averaging success rates over all runs. Thus, no model is tested on compounds that were part of its training data.

4 Experimental Setup

Recall that for the actual translation of a given compound, its component parts are looked up in the bi-lingual dictionary, and each component translated into its English counterparts. We will refer to the fertility of each component as n_1 and n_2, where for our example (1-a) above, say, $n_1 = 22$ and $n_2 = 5$, i.e. there are 22 available translations for the noun *anlegg* and 5 for *vei*, respectively.

4.1 Preparatory Steps

All component translations are 'slotted' into the translation templates, resulting in a set of translation candidates. The total number of candidates is the cross-product of n_1, n_2, and the number of distinct templates (20, in our experiments). This is indeed one of the richer examples, and in our experiments the maximum number of translation candidates did not exceed a couple of thousand possible outcomes. For each translation candidate, a set of quantitative corpus data is extracted from the pre-processed and indexed TL corpus. These data are then used to rank the candidates, in various ways, either by means of the CTQ of Baldwin and Tanaka (2004), or as the input to the MaxEnt ranker. While in the former (heuristic) case the corpus data can be directly used for ranking and testing on the gold standard (there is no separate training step), the MaxEnt approach requires separate training and test data sets, which we address by ten-fold cross-validation over the gold standard.

The splitting up of compounds (using the optional OBT component mentioned earlier) and component translation was carried out as a preparational step, where each SL compound and its component parts with TL translations were indexed in an intermediate data structure.

4.2 Candidate Generation with Templates

It was a requirement in the implementation that the Norwegian compounds could be split up into two parts, both of which were nouns. For the English translation, however, it is accepted that one of the components be translated as multiple English words, as in example (2). To accommodate this variation, all TL frequency counts discussed below can in principle range over any TL phrase, as observed in any of the candidate translations are any of the 'slots' defined by our set of translation templates.

(2) hytte·tilsyn
cottage·supervision agency
'cottage supervision agency'

4.3 Ranking Baseline: Reference

For the ranking task, as a simple baseline (i.e. a measure of how the more refined ranking methods performed), a reference ranking based on only the frequency (in the available TL corpora) of the translation candidate *in full* was introduced. Of two candidates, such as 'down bag' vs. 'bag of down', the most frequent phrase would be chosen.

4.4 Corpus-based Translation Quality

A much stronger baseline, borrowed from Baldwin and Tanaka (2004), was used—the interpolated CTQ metric[7]—which extracts the frequency

[7]Baldwin and Tanaka (2004) give a slightly revised formalization for CTQ, as compared to the earlier version of Tanaka and Baldwin (2003b). Furthermore, in the earlier publication there is room for uncertainty as to whether each term, esti-

Mono-Lingual Features
CTQ
freq(E_1, E_2, t)
freq$(E_1, _, t))$
freq$(_, E_2, t)$
freq(E_1, t)
freq(E_2, t)

Table 1: Corpus-based MaxEnt features, where E_1 and E_2 denote English phrases 'slotted' in as the first or second element of a compound template t. Most often, E_1 and E_2 are single words.

counts from the target language corpus.

$$\text{CTQ}(w_1^E, w_2^E, t) = \\ \alpha p(w_1^E, w_2^E, t) + \beta p(w_1^E, t) p(w_2^E, t) p(t) \quad (1)$$

Equation 1, firstly computes the probability of two English words, w_1 and w_2 occurring as an instance of the template t, multiplied by an interpolating weight, α, then adds the product of the probability of w_1 as the first element in a construction licensed by template t and the probability of w_2 being the second element, respectively. An example would be the count of *machine translation* occurring as two nouns in a sequence (the template) divided by the total count of all template instances, added to how often *machine* is the first word of such couples, and *translation* is the second, to capture what words more often let themselves be combined in such compounds.

4.5 MaxEnt Basics: Mono-Lingual Features

The Maximum Entropy (MaxEnt) framework has been applied successfully to NLP tasks before (Ratnaparkhi, 1996; Ratnaparkhi, 1998; Mikheev, 2000; Charniak and Johnson, 2005; Velldal, 2008) in areas like parsing, sentence boundary detection, and PoS tagging, but notably (re-)ranking, for which it is also used in this paper. The various statistics for each translation candidate (which will be discussed in further detail below), can be used as features in a conditional MaxEnt model (the family of MaxEnt models is also commonly referred to as log-linear or exponential models).[8]

mated by maximum likehood over the training corpus, should be conditioned on t or not: Tanaka and Baldwin (2003b) discuss the terms as 'conditional' probabilities, but equation 1 suggests a non-conditional formalization (in contrast to, for example, $p(w_1^E, w_2^E|t)$). We implemented both variants and found the non-conditional CTQ to perform substantially better, hence restrict ourselves to this variant in the following. Just like T&B, we use $\alpha = 0.9$ and $\beta = 0.1$.

[8]Like Velldal (2008) and much other current work, we make use of the open-source TADM framework, see `http://tadm.sourceforge.net` (Malouf, 2002).

Bi-Lingual Features
freq$(E_1, E2|N_1, N2)$
freq$(N_1, N2|E_1, E2)$
freq$(E_1, E2, \rightarrow)$
freq$(E_1, E2, \leftarrow)$
freq$(E_1|N_1)$
freq$(E_2|N_2)$
freq$(N_1|E_1)$
freq$(N_2|E_2)$

Table 2: Bi-lingual features, extracted from the dictionary. N_1 and N_2 denote the first and second element of the Norwegian compound and E_1 and E_2 designate the English translations of these components in the current translation template.

Given a source language compound n, our model estimates the probability of a candidate translation e_i as the normalized dot product of a vector $\vec{f}$ of so-called features—arbitrary properties determined by so-called feature functions—and a vector $\vec{\lambda}$ of corresponding weights:

$$p(e_i|n) = \frac{\exp \sum_j \lambda_j f_j(e_i, n)}{\sum_{k=1}^{n} \exp \sum_j \lambda_j f_j(e_k, n)} \quad (2)$$

The search for the highest-scoring candidate can then be formalized as $\arg\max_{e_i} p(e_i|n)$, i.e. finding the translation candidate e_i that maximizes the conditional probability, given n. The machine learning task, then, is to find the vector $\vec{\lambda}$ that maximizes the (conditional) likelihood of the training distribution—a problem for which off-the-shelf solutions are available.

To avoid the stipulation of linear interpolation weights in CTQ, we defined a MaxEnt model with a feature set consisting solely of (log-)frequencies extracted from the target language corpus. For all MaxEnt models that were built, an additional binary feature identifying the template, which would inform the model on which template was the most frequent, was used. The mono-lingual features that were used are shown in Table 1.

4.5.1 MaxEnt with Bi-Lingual Features

In addition to the two experiments testing the difference between humanly estimated interpolation weights and the results of using a machine learning engine, the MaxEnt learner was also tested on a full feature set, with features also encoding information about the individual translation(s) of the source input, and not just the monolingual target language features of the translation candidate. Our bi-lingual feature set, extracted from the one Norwegian – English dictio-

nary available, is summarized in Table 2. In this model, bi-lingual features are added 'on top' of the mono-lingual ones.

These dictionary-based features indicate how often an English component E_1 or phrase E_1E_2 is counted as a translation of its Norwegian source. Because there can be multiple senses of an entry in the dictionary, a translation can have frequencies above 1, meant to capture what is a more likely translation for a given source word. In addition, frequencies of the translation candidates attested in the dictionary, regardless of the source are captured, as well as using the dictionary in both directions. In Table 2 the symbol '$\rightarrow$' indicates use of the dictionary in 'forward' direction (Norwegian – English), and '$\leftarrow$' the reverse direction.

4.6 Variation in Analysis Depth

The RASP analyzer was used for the preprocessing of the English language text corpora. RASP results were then searched by means of regular expressions, corresponding to the TL side of our translation templates, in order to extract the frequency of the various types of translations. In performing these queries, there is a choice as to whether to use RASP annotations only at the part-of-speech (PoS) level, or whether to inspect full phrase chunks. Consider the simplified examples (3) and (4), showing attachment of a 'for' PP either inside of an NP, or as a VP modifier instead:

(3) (VP (VB buy)
(NP (NNS books)
(PP (IN for) (NP (NN children)))))

(4) (VP (VB buy) (NP (NNS books))
(PP (IN for) (NP (NN children))))

If the regular expression used for counting occurrences of the $\langle E_2$ for $E_1\rangle$ template only inspected the PoS tags associated to each word, both (3) and (4) would match, resulting in a false positive count. A regular expression query requiring all template elements to be embedded inside an NP, on the other hand, would count only the first one. Seeing that RASP annotations are fully automated, where the syntactic layer is bound to have a higher error rate than the PoS layer, however, it is not *a priori* known which of the two strategies would yield better approximations of the actual counts. Variation of analysis depth, in this sense, is a dimension of variation to all experiments summarized in Section 5 below.

4.7 Variation in Corpus Size

The experiments were conducted using the corpora BNC, AQ and NAN (as mentioned in Section 3), where additional training data was added incrementally, starting with only the BNC, then adding AQ, and finally also adding NAN. The amount of training data used is another, orthogonal dimension of variation to the experimental results reported below.

4.8 Parameter Tuning — Implementation

The TADM MaxEnt toolkit allows the tuning of certain hyper-parameters to the estimation process. Feature weights can be smoothed using a so-called Gaussian prior, and relative or absolute tolerance thresholds can be applied in determining learner convergence. A large space of different combinations for these hyper-parameters was explored experimentally, but learner performance was relatively stable within substantial intervals around the TADM default values; no specific combination lead to significantly improved performance, when compared to the default hyper-parameters. Thus, all results reported here assume standard TADM settings.

5 Results

An overview of experimental results can be found in Table 3, where REF denotes the simple frequency baseline, CTQ the original T&B metric, ME_1 our mono-lingual MaxEnt model, and ME_2 the full MaxEnt model, including dictionary features. The results show a notable increase in performance as we go from REF- and CTQ-based ranking to MaxEnt ranking, and a smaller, yet significant increase as the bi-lingual features are introduced. The increase between REF and CTQ shows how the weighted information about the 'association strength' between single component corpus data and the translation candidate itself boosts performance; and the difference between CTQ and ME_1 shows that it helps to combine these data through a principled machine learning approach. The fully superior performance of the MaxEnt model with all features, finally, suggests that adding more information (by way of features) to the model increases performance further.

In the following few paragraphs, we discuss these results further, along the various dimensions of variation that we have set up for these experiments.

		REF		CTQ		ME_1		ME_2	
Corpora	Band	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*	*Tagger*	*Parser*
BNC	high	28.03	25.00	32.58	31.82	39.80	38.70	51.10	51.90
	middle	20.51	19.23	26.28	33.33	31.80	36.30	51.20	50.90
	low	12.10	11.46	19.11	24.20	33.60	31.80	45.90	48.90
	all	19.77	18.20	25.62	29.65	34.81	35.42	49.3	50.49
+AQ	high	38.64	35.61	40.91	40.91	49.80	54.90	57.40	59.70
	middle	23.72	25.00	30.77	36.54	39.00	41.10	52.00	54.20
	low	13.38	12.10	19.11	20.38	26.80	27.80	45.50	46.70
	all	24.50	23.59	29.66	32.12	37.90	40.5	51.31	53.18
+NAN	high	35.61	37.12	38.64	38.64	49.40	51.60	58.70	59.60
	middle	23.08	24.52	26.92	29.03	38.60	39.60	51.80	52.20
	low	16.56	14.01	18.47	17.20	25.80	26.50	48.00	45.50
	all	24.50	24.55	27.42	27.70	37.28	38.54	52.51	52.03

Table 3: Overview of gold standard results, measured as the percentage of correctly translated compounds.

Frequency Bands In the success figures of Table 3, there is a general tendency across ranking methods to perform better on high-frequency compounds, presumably because frequency of use will impact the reliability of statistics used in ranking. We have not investigated this effect in a systematic manner, but recall from Section 3 that (a) the frequency bands were established from web counts (we lack a Norwegian corpus of sufficient size) and (b) our compound discovery procedure using the Oslo-Bergen Tagger is biased, in that a large number of compositional but frequent compounds have been entered into the system lexicon (as simplex words) and, hence, are omitted from our study. Thus, results presented here probably under-estimate the actual performance of our method.

Analysis Depth Table 4 shows the differences in performance between using tagger-based and parser-based data. For the three ranking methods displayed in the table, the parser-based generally data show an improvement in performance, i.e. the added precision of counts taking into account syntactic structure seems to outweigh the expectation of a higher error rate in RASP results at this higher depth of analysis. For all ranking methods, however, the difference is smallest when all training corpora are used, and parser-based counts even yield a slightly lower performance for the full corpus using all MaxEnt features (i.e. our most advanced model).

Corpus Size As Table 3 indicates, the performance of the various rankers generally increases as the base corpus from which quantitative data are extracted is larger. But it is also evident that going from the BNC to the BNC+AQ combination shows the biggest difference in performance. In fact, going from there to +NAN surprisingly indicates a decrease in performance, except for one set of experiments. The difference, however, is very small for the the most sophisticated ranking method, the fully-featured MaxEnt model. For 38 Norwegian compounds the top-ranked translation candidate diverged for the +AQ and +NAN experiments, with half of them going in either direction. Hence, a sign test exploring the likeliness of this result if the two methods +AQ and +NAN are equal, would find such an outcome expected, if the 'methods' are equal.

6 Discussion

Our experiments show that the MaxEnt approach is viable to finding the correct translation of nominal compounds, just as Baldwin and Tanaka (2004) show how a SVM can give better results than humanly stipulated interpolation weights. The performance also increases as a full feature set is used, including translation counts for the individual compound subparts, instead of only frequencies of the translation candidate itself.

The MaxEnt approach allows just for this combination of features, both features stemming from linguistic insight, as well as purely quantitative measures resulting from counts from annotated corpora. It will be possible to introduce further semantic information into such a model, when available, depending on the framework in which it is implemented. In our experiments, only one bilingual dictionary was used (Eek, 2001), but the

Corpora	REF	CTQ	ME_1	ME_2
BNC	-1,65	3,80	0,53	1,17
+AQ	-1,01	2,35	2,73	1,9
+NAN	0,14	0,28	1,3	-0,4

Table 4: Difference in performance when RASP is used as a parser and a tagger. A negative figure shows that tagger-based counts led to better ranking results.

counts for a translation could vary because of the different senses of one word stored in a lexicon entry. There may, however, also be other systematic relations between a compound and its correct translation, for example a relationship between a certain joint element and the output construction type, or the between semantic information and construction type. Such features could be implemented through the use of binary features, allowing them to be included in a MaxEnt model.

Although a larger corpus would likely yield better coverage of rare constructs, and accordingly help overall performance, a decrease in marginal benefit from adding words would also be expected. The low frequency band benefits less from the enlargement of the corpus, whereas the middle and high frequency bands show a marked improvement going from BNC to BNC+ANC. Our expectation was that the lower frequency band would benefit more from better coverage in the basis corpus, so this was an unexpected result. More research is needed to verify or explain this tendency.

References

Baldwin, Timothy and Takaaki Tanaka. 2004. Translation by Machine of Complex Nominals: Getting it right. In *Proceedings of the ACL04 Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain.

Briscoe, Ted, John Carroll, and Rebecca Watson. 2006. The Second Release of the Rasp System. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, Australia.

Charniak, Eugene and Mark Johnson. 2005. Coarse-to-fine n-best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting of the ACL (ACL05)*, pages 173–180, Ann Arbor, MI, USA.

Eek, Øystein, editor. 2001. *Engelsk stor ordbok: engelsk – norsk/norsk – engelsk ('English Large Dictionary')*. Kunnskapsforlaget, Oslo, Norway.

Grefenstette, Gregory. 1999. The World Wide Web as a Resource for Example-Based Machine Translation Tasks. In *Translating and the Computer 21: Proceedings of the 21st International Conference on Translating and the Computer*, London, UK.

Hagen, Kristin, Janne Bondi Johannessen, and Anders Nøklestad. 2000. A Constraint-based Tagger for Norwegian. In *17th Scandinavian Conference of Linguistics*, Odense, Denmark.

Johannessen, Janne Bondi and Helge Hauglin. 1996. An automatic analysis of norwegian compounds. In *Papers from the 16th Scandinavian Conference of Linguistics.*, Turku, Finland.

Malouf, Rob. 2002. A Comparison of Algorithms for Maximum Entropy Parameter Estimation. In *Sixth Conf. on Natural Language Learning*, pages 49–55, Taipei, Taiwan.

Mikheev, Andrei. 2000. Tagging Sentence Boundaries. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 264–271, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Oepen, Stephan, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, and Victoria Rosén. 2004. Som å kapp-ete med trollet? Towards MRS-based Norwegian – English Machine Translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, USA.

Rackow, Ulrike, Ido Dagan, and Ulrike Schwall. 1992. Automatic Translation of Noun Compounds. In *Proceedings of the 14th Conference on Computational Linguistics*, pages 1249–1253, Nantes, France.

Ratnaparkhi, Adwait. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In Brill, Eric and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey, USA.

Ratnaparkhi, Adwait. 1998. Maximum Entropy Models for Natural Language Ambiguity Resolution. Technical report, University of Pennsylvania.

Tanaka, Takaaki and Timothy Baldwin. 2003a. Noun-noun Compound Machine Translation: A Feasibility Study on Shallow Processing. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

Tanaka, Takaaki and Timothy Baldwin. 2003b. Translation Selection for Japanese-English Noun-Noun Compounds. In *In Proceedings of Machine Translation Summit IX*, New Orleans, LO, USA.

Velldal, Erik. 2008. *Empirical Realization Ranking*. University of Oslo, Oslo, Norway.