

Course Assignment in Natural Language Processing, GSLT

On the automatic translation of noun compounds: challenges and strategies

Ebba Gustavii

22nd May 2004

1 Introduction

In this report I will discuss the challenges involved in automatic translation of compound words and give a brief review of strategies employed for addressing some of these issues. More specifically, I will be concerned with the translation of productively formed noun+noun compounds.

Several aspects of compounds make them particularly difficult to handle in a system performing automatic translation, being it a machine translation system or a system for cross-language information retrieval. First, the relation between the parts of a compound is implicit and thus its interpretation is never wholly compositional. Whereas the interpretation of clauses is guided by syntactic clues, such as word order and morphemic markers, the meaning of a compound cannot be fully recovered from the surface structure. This is, in particular, a problem when translating from a language with frequent use of compounds, to a language generally preferring syntactic constructions instead, since an overt syntactic marker (usually a preposition) is then to be generated. Secondly, also languages with frequent use of compounds differ as to when the use of a compound is preferred over some other construction type - that is, it seems in part to be an arbitrary decision of each language when to compound.

The term compound noun will here be used to denote a fused or juxtaposed unit of a head noun and a modifying noun, for which there are no overt syntactic markers indicating the relation between the parts. Further, we will only be concerned with endocentric compounds, i.e. compounds where one of the parts functions as the semantic and syntactic head of the compound as a whole.

The remainder of this paper is organised as follows. In section 2, the challenges involved in translating compounds will be discussed. In section 3, I will summarise four approaches taken to some of the involved issues, two of which are interpretation-driven and two that are based on shallow processing.

2 Challenges

Depending on the source and target languages, the challenges involved in translating compounds are somewhat different. When translating from languages such as German and Swedish, where the compound parts are fused, the first task is to properly divide the compound into simplex words. Often there is a multitude of possible segmentations, and the main difficulty lies in picking the correct one. For this task, several successful strategies have been suggested, see for instance (Dura 1998). In general though, these strategies are targeted at picking the most plausible segmentation of a compound regardless of its context, such as generally preferring *publik+underlag* over *pub+lik+under+lag*. Less is done on resolving segmentation ambiguities that are contextually dependent, as in¹:

Kulturgeshichte Kultur+Geschichte → *history of culture*
Kult+Urgeschichte → *pre-history of worship*

A second problem, present also in languages with juxtaposed compound parts, lies in establishing the correct hierarchical structure when the compound includes more than two root morphemes. Consider, for instance, the Swedish compound (*jordbruks+stöds*)+*process*) and its English translation *procedure for agricultural aid*². The translation implies the analysis where *process* is modified by *jordbruksstöd*, and would not be appropriate otherwise.

The problems of segmentation and bracketing are not specific to machine translation, but are addressed in connection with a range of NLP-applications. I will therefore not go into them any further here. Neither will I specifically go into the question of target word selection (as regards lexical content), since that is one of the main issues focused on in machine translation in general.

Assuming then that we have established the correct segmentation, bracketing and target lexemes, the remaining tasks may be summarised as follows³:

- Selection of target construction type
- Choice of syntactic marker (when target structure is a phrase)
- Choice of morpho-syntactic features (when target structure is a phrase)
- Integration of target structure with the rest of the sentence

Since languages differ as to when to compound, a source language compound is not always most properly translated as a compound. For each source language compound, the appropriate target construction must thus be determined. Consider for instance the following Swedish compounds and their respective English translations²:

¹Example taken from (Hutchins and Somers 1992)

²Examples taken from a parallel corpus of agricultural EC-texts.

³The list is in part taken from (Pease 1993)

<i>resurscentrum</i>	resurs+centrum	→	<i>resource centre</i>
<i>naturresurs</i>	natur+resurs	→	<i>natural resource</i>
<i>resursbrist</i>	resurs+brist	→	<i>lack of resources</i>

In the first example the source construction type is preserved. In the second example, the modifying noun is translated as a relational adjective i.e. an adjective semantically related to a noun (in general also derivationally) which is not gradable nor used predictively (Bennet 2002). This construction type is so similar to noun compounds, both in terms of function and distribution, that some researchers consider them as being compounds rather than phrases (Paggio and Ørsnes 1994) (Bennet 2002). The final example illustrates the translation of a modifying noun to a prepositional post-attribute. It should be noted though, that there often are several legitimate construction types to choose from, something which naturally follows from the fact that compounds generally may be paraphrased.

When the chosen target construction type involves a prepositional post-attribute, the appropriate preposition must be selected, as may be illustrated by the following German-English examples taken from (Copestake and Lascarides 1997):

<i>Terminsvorschlag</i>	Termin+Vorschlag	→	<i>proposal for a date</i>
<i>Terminsvereinbarung</i>	Termin+Vereinbarung	→	<i>agreement on a date</i>

When translating a modifying noun as a prepositional attribute the appropriate number and definiteness features of the governed noun must also be selected. In the compound context, the modifying noun is uninflected and is generally interpreted with a conceptual, generic meaning (Teleman, Hellberg and Andersson 1995). According to (Pease 1993), this meaning is generally best transferred as the undetermined plural, as in *idéutbyte* → *exchange of ideas*. There are however frequent exceptions, sometimes due to the generic meaning being realised as the singular *bilresa* → *journey by car*, and sometimes due to the modifying noun having a contextually dependent referential meaning *artikelförfattare* → *author of the articles/article*.

Number and definiteness of the head noun are usually preserved in the translation, though there are interesting exceptions. The following example, inspired by (Pease 1993), illustrates how the relation between the compound parts influences the selection of definiteness of the translated phrase (the crown is always unique in relation to a tree): *en trädkrona* → *the crown of a tree*. The indefiniteness of the source language phrase is not signaled on the overall phrase as is usual, but rather on the modifying noun. This comes clear when we translate the definite counterpart: *trädkronan* → *the crown of the tree*.

Integrating the translated compound into its context may also present difficult choices. One specific issue concerns potential attributes to the compound. Generally, these modify the compound as a whole, and do not operate only on the non-head. This explains why we cannot denote an apple tree with tasty apples by *a tasty apple tree*. But sometimes this restriction is overridden, as is evident from

the English translation of *seinem Lebensende* as *the end of his life* and not as *his end of life* (Pease 1993).

3 Approaches

There are different views on the level of analysis required for proper translation of noun-noun compounds, and correspondingly, the proposed strategies may generally be classified as either deep or shallow. In strategies based on deep processing, there is an attempt to semantically analyse the compounds, and the translation relation is defined via an intermediate representation. With shallow approaches, the translation is based directly on the source language words.

3.1 Interpretation-based approaches

Quite naturally, interpretation-based strategies have mainly been proposed in relation to translation tasks from languages with frequent use of compounds to languages where other construction types are generally preferred (Paggio and Ørsnes 1994) (Gawronska, Nordner, Johansson and Willners 1994) (Navigli, Velardi and Gangemi 2003). Under such settings, the problem of generating a syntactic marker indicating the relation between the parts becomes more emergent.

Though lots of research has been done on the interpretation of compounds, not so much has been targeted at establishing the importance of such an interpretation for the automatic translation of compounds.

(Paggio and Ørsnes 1994) present a study on the feasibility of semantic analysis for translating Danish nominal compounds into Italian. As is common when dealing with compound interpretation, they distinguish between argumental and non-argumental compounds. The head of an argumental compound is generally derived from a verb for which the non-head fills an argumental slot e.g. *arbejdsdeling* (Danish for *division of work*). According to Paggio and Ørsnes, the analysis of these will not impose major difficulties, since they claim there to be clear preferences for which argument will fill the non-head position, and that the preposition to generate will be determined by the valency restrictions of the Italian head-noun. Instead they focus on non-argumental compounds. There have been different opinions on the type of relations that may hold between the non-head and a head in such compounds. On the one extreme, (Levi 1978) claims that there is a finite number of possible relations and proposes a list of those. (Selkirk 1982) on the other hand, defies the possibility of ever capturing the broad range of relations. Paggio and Ørsnes follow the stand taken by Levi, and classify a set of 287 Danish compounds as having one out of 12 possible relations. For instance, the FROM relation is assigned to *forskningsresultat*. Having classified the Danish compounds, they look for regularities in how they are translated into Italian. The distribution of translational construction types turns out to be rather unpredictable from the basis of the semantic relation holding between the head and the non-head; only two rather

unambiguous mappings are established. These mappings, together with a default Italian construction type, are however enough to predict the correct construction type for 88.8% of the compound set.

(Navigli et al. 2003) also experiments with an interpretation-driven approach to the translation of English compounds to Italian, though they narrow the task to involve only complex terms from a limited domain. By using a machine-learned ontology, the parts of the source language compound are first disambiguated and mapped to concepts (WordNet synsets). To assign the correct relation holding between the compound parts (out of 10 possible), an inductive machine learning technique is used. The learned rules are based on triplets of concept-relation-concept, where each concept is represented by its set of WordNet hypernyms. Italian lexemes are then generated using EuroWordNet, and the conceptual relation is mapped to an Italian preposition by hand-crafted rules. Since there is not always a single mapping from a conceptual relation to a preposition, several suggestions will often be generated. The same of course holds for the synset translations. To select the best out of the generated candidates, Google is searched and the one with the most hits is chosen.

Summing up, the appropriate construction for Italian does not seem to be directly derivable from the intermediate semantic representations suggested here. In the first example only two, out of the twelve, possible relations gave unambiguous clues, and in the second example the interpretation-driven strategy had to be supplemented with target language data (as represented by the web-search). Of course, these findings may have to do with the specific set of parameters used - not least the chosen set of semantic relations. It is further not obvious how the results relate to generation in other languages.

The reported approaches only target the problem of finding the appropriate construction type as regards the choice of preposition (and to some extent also the possibility of choosing a relational adjective). They do not address the problems involved in the appropriate number and definiteness of the non-head nor any integrational issues. Further, they work on compound types (as opposed to tokens in context) assuming there is only one possible relation between the parts (something which of course makes more sense in the second example where targeted compounds are terms).

3.2 Shallow approaches

Several strategies based on shallow processing have been proposed for the translation of compounds. The simplest ones rely on static transfer dictionaries that list compounds, either obtained from aligned parallel corpora or by traditional lexicographic work. Though these suffer from the obvious shortcoming of not being able to cope with the productive formation of new compounds, they are commonly used. (Tanaka and Baldwin 2003a) More elaborate strategies dynamically assign translations to source language compounds. These strategies typically rely on statistical data collected from large corpora.

Since a content expressed as a compound in one language may best be expressed by some other construction type in another language, the use of compounds cannot be solely explained in terms of semantic and pragmatic factors, but is in part a question of language-specific usage. As (Rackow, Dagan and Schwall 1992) put it, it thus “seems suitable to look for the information [of construction type] in a target language corpus”. Though example-based machine translation generally uses bilingual (parallel) data, the strategies proposed for compound translation are mainly based on the use of mono-lingual corpora. These are easier to obtain in very large sizes thus alleviating the problem of data sparseness.

3.2.1 Word-level dictionary and target corpus

(Tanaka and Baldwin 2003b) proposes an approach based on a word-level translation dictionary and a target language corpus. Though their experiments are targeted at translation from Japanese to English the strategy is portable to other language pairs (similar work has been done on the translation of German compounds to English (Rackow et al. 1992)). To translate a productively formed noun+noun compound, they combine the possible word-level translations with a set of constructional translation templates. The latter define possible structural mappings from source language compounds to the target language. The following template, for instance, defines a mapping to an English noun with a prepositional post-attribute: $[N_2^E \text{ in } N_1^E]$ (where N_i^E denotes an English (E) noun (N) corresponding to the i -th source language lexeme). 28 such templates were established on the basis of a set of validated translation-pairs. By taking all possible combinations of the word-level translations, and slotting these into the templates, a set of potentially legitimate translations is generated. For selecting the most plausible one out of these, empirical data from a target language corpus is used. Each candidate is scored by interpolating probabilities⁴ of fully and partially specified translation data (where the template itself counts as a piece of translation data). The strategy was evaluated on a test set of 500 Japanese noun+noun compounds, for which a gold standard was first established. Since a compound may have more than one legitimate translation, there were often sets of translations considered acceptable. They report an F-score of 0.68.

The use of non-aligned data may of course give misleading clues as to how the content of a compound is best generated. The choice of construction type is sometimes guided by factors such as the presence of modifying attributes (as in the example with the apple tree). Tanaka and Baldwin retrieves the target language data from a dependency-parsed corpus, and count instances of construction types regardless of there potentially being intervening attributes. Thus a content, which by itself best would be expressed as a compound, may be represented by phrasal constructions in the corpus.

⁴Calculated according to a maximum likelihood estimate.

3.2.2 Contextual similarity

Following the strategy proposed in (Tanaka 2002), the best translation candidate is not assumed to correspond to the most frequently occurring candidate in a target language corpus, but rather to the candidate that appears in the contexts most similar to that of the source language compound. In practice, this strategy provides a modification of the selection phase in the previous strategy. Sets of translation candidates are first generated by translating the source language lexemes word-by-word, slotting these into templates and then keeping only those combinations that are attested in a target language corpus. Note that the strategy thus only accounts for translations that are attested as wholes, and that it thereby will have difficulties translating completely novel compounds. Word translations are quite generously defined; source and target lexemes need not be linked via a dictionary, but may be linked via thesauri. This is motivated by compound translations such as *capital investment* from the Japanese compound *setsubi toushi*, the direct translation of which would be *equipment/facility investment*.

To select the most plausible translation, the context of source language instances in a mono-lingual corpus is compared to those of the candidate translations in a target language corpus. Context is defined by words co-occurring in the same sentence, and is represented by vectors where each position defines the relative relevance of a word given the compound⁵. Source language vectors are converted to target language equivalents by multiplication with a predefined translation matrix. An evaluation of the selection method shows that the best translation candidate was chosen in more than 70% of the cases. In comparison to a baseline based on frequencies, the results proved to be significantly better. Similar approaches have been taken by (Shahzad, Kiyonori, Shigeru and Kazuhide 1999) and (Cao and Li 2002) (the latter for the translation of Chinese compounds to English).

4 Concluding remarks

There are several challenges involved in the automatic translation of compounds, and there are also quite different approaches suggested to deal with some of these issues. Though only a few strategies have been presented here, I believe those to be quite representative in a number of ways. First, they illustrate the main division between approaches based on deep and shallow processing; a division common to machine translation strategies in general. The method proposed by Navigli et al illustrates a tendency towards blending features from both paradigms. Secondly, they all seem to ignore the fact that the relation holding between head and non-head, at least to some extent, is context dependent.⁶ This may be taken as an indication of either the strategies simplifying matters, or that in practice, the context is only

⁵The relative relevance of the co-occurring words is calculated by a log-likelihood ratio as defined by (Dunning 1993).

⁶This is true also for the strategy based on contextual similarity since all instances of a source language compound are treated the same.

minimally influencing the relation between compound parts. Thirdly, the strategies are mainly addressing target construction type, both regarding the question of compounding versus non-compounding, and the choice of syntactic marker. Issues involving the inflection of the non-head, and the integration of the target construction in its context, are left for future research.

References

- Bennet, P. (2002). English Adjective-Noun Compounds and Related Constructs, *GEMA: Online Journal of Language Studies*.
- Cao, Y. and Li, H. (2002). Base noun translation using Web data and the EM algorithm, *COLING-2002*, Taipei, Taiwan.
- Copestake, A. and Lascarides, A. (1997). Integrating symbolic and statistical representations: The lexicon pragmatics interface, *ACL-EACL*, Madrid, Spain.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*.
- Dura, E. (1998). *Parsing Words*, Novum Grafiska, Göteborg.
- Gawronska, B., Nordner, A., Johansson, C. and Willners, C. (1994). Interpreting compounds for machine translation, *COLING-1994*, Kyoto, Japan.
- Hutchins, J. and Somers, H. (1992). *An Introduction to Machine Translation*, Academic Press Ltd, London.
- Levi, J. (1978). *The Syntax and Semantics of Complex Nominals*, Academic Press, New York.
- Navigli, R., Velardi, P. and Gangemi, A. (2003). Ontology Learning and its application to automated terminology translation, *IEEE Intelligent Systems*.
- Paggio, P. and Ørsnes, B. (1994). Automatic Translation of Nominal Compounds - A Case Study of Danish and Italian, *Rivista di Linguistica*, Vol. 5:1, Rosenberg & Sellier, Torino.
- Pease, C. (1993). The analysis of German compounds and their translation into English. Report in the cat2 project.
- Rackow, U., Dagan, I. and Schwall, U. (1992). Automatic translation of noun compounds, *COLING-1992*, Nantes, France.
- Selkirk, E. (1982). *The Syntax of Words*, MIT Press, Cambridge, Mass.
- Shahzad, I., Kiyonori, O., Shigeru, M. and Kazuhide, Y. (1999). Identifying Translations of Compound Nouns Using Non-aligned Corpora, *Proceedings of the workshop MAL-99*.

- Tanaka, T. (2002). Measuring the similarity between compound nouns in different languages using non-parallel corpora, *COLING-2002*, Taipei, Taiwan.
- Tanaka, T. and Baldwin, T. (2003a). Noun-noun compound machine translation: A feasibility study on shallow processing, *ACL*, Sapparo, Japan.
- Tanaka, T. and Baldwin, T. (2003b). Translation Selection for Japanese-English Noun-Noun Compounds, *MTSUMMIT-IX*, New Orleans.
- Teleman, U., Hellberg, S. and Andersson, E. (1995). *Svenska Akademiens Grammatik*, NorstedtsOrdbok.