AUTOMATIC TRANSLATION OF NOUN COMPOUNDS

ULRIKE RACKOW
IBM Scientific Center
Institute for Knowledge
Based Systems
Heidelberg, Germany
rackow@dhdibm1.bitnet

IDO DAGAN
Computer Science Department
Technion, Haifa, Israel *and*
IBM Advanced Solution Center
Haifa, Israel
dagan@cs.technion.ac.il

ULRIKE SCHWALL
IBM Scientific Center
Institute for Knowledge
Based Systems
Heidelberg, Germany
schwall@dhdibm1.bitnet

Abstract

This paper describes the treatment of nominal compounds in a transfer based machine translation system; it presents a new approach for resolving ambiguities in compound segmentation and constituent structure selection using a combination of linguistic rules and statistical data. An introduction to the general as well as to the German-English-specific problems of compound translation is given (sect. 1). In section 2, the analysis phase is described with its linguistics as well as its computational aspects. Section 3 deals with the transfer and generation process, focussing on corpus based techniques.

1 Introduction

It is widely known that the word formation mechanism of compounding is highly productive, in German as well as in English, and that efficient strategies have to be developed to deal with this linguistic phenomenon in any kind of NLP system. Although this fact is generally agreed upon and a lot of linguistic research has been done, it has not been possible so far to develop a general and overall procedure to solve the problem in a satisfactory and adequate way (cf. [Ananiadou/McNaught 1990]).

Two special aspects of the problem of compounding phenomena arise, among others, within the framework of machine translation (MT), here the translation from German into English. The first problem that has to be dealt with in this case is the correct segmentation of the German compound word. The constituents having been found, the next step we have to deal with consists in translating them correctly. Correctness refers here a) to the choice of the appropriate target lexemes and b) to the selection of the right target construction type.

Of course, there are a lot of other problems to be resolved for the treatment of compounds in MT, e.g. semantic interpretation of the relation between the constituents, the question in how far this point is relevant for translation, depth of analysis, etc. In this paper, however, we will mainly concentrate on the two problems mentioned above.

An important property of our approach for segmentation (cf. 2) is optimizing the process by using the type of the juncture between the compound constituents to formulate restrictions on their possible position (front, middle and/or end) in the compound word. Another novel characteristic of our approach is that there is no need of finding out the correct constituent structure during analysis phase. This problem is transferred to the process of selecting the correct target compound construction (cf. 3.3). The solutions we propose are based on an investigation of examples which were extracted, in part randomly, from real text corpora.[1] Contrary to the approach of example-based machine translation (e.g. cf. [Sumita 1991]), we don't use a bilingual corpus, but a monolingual target corpus which is much easier to obtain in a very large size. The last feature of our approach we would like to point out here is its multilinguality: on the one hand, the results of German compound analysis can serve as input for all target languages; and, on the other hand, the features of the English construction types associated with the target entries for English nouns can also be used for source languages other than German, and what is important, for NLP-applications other than MT.

The several components of our model are currently being tested separately, and an integration is planned. Preliminary results indicate that the corpus based techniques achieve high accuracy, but they are not fully analyzed yet. We plan to report complete results in a future paper.

2 Automatic Analysis of German Compounds

2.1 Preliminary Remarks

Our work focuses on nominal compounds; in our first approach, we narrowed this type even more to *noun-noun(-noun...)* compounds, these constructions being again the most frequent type of nominal compounds in both languages (cf. [Rackow 1992]). This restriction to nouns gives us the possibility of using the part-of-speech in the segmentation algorithm to reduce the number of possible segmentation results; in any case, certain personal or possessive pronouns, conjunctions etc. can be excluded explicitly for they never occur in productive composition types. This way, we can avoid wrong decompositions, such as *Uns-Innigkeits-Vorwurf ('us-intimacy reproach') instead of Unsinnigkeits-vorwurf ('nonsense reproach').

Only those compounds which are not lexicalized are treated, i.e. the segmentation and translation algorithm is only called upon if an input word has not

[1] The German examples are partly taken from the SPRING-Corpus which was kindly put at our disposal by the Speech Recognition Group of the German IBM Science Center Heidelberg. The English data were extracted from the corpora maintained by the speech group of IBM Watson Research Center, Yorktown Heights.

been found in the system's lexicon.

With German as the source language, the first problem in the treatment of compound words arises from the fact that German compounds are written in one word and that in many cases, the form of the words in a compound differs from the base form in that either a so-called *Fugenelement* (connecting element or juncture morpheme) is added to the modifying word or that one or more letters are taken away from the ending of these words. In order to allow for a correct segmentation of the compounds, a code has to be added next to the morphological declension code of the entries in the analysis part of the lexicon pointing to the corresponding morpheme types.

2.2 Code for the Connecting Element

The importance of the correct encoding of the connecting element is shown in the following example. Suppose a word like *Arbeitsamt* — *'job center'* would not have an entry in the lexicon and *Arbeit* would not be encoded with the connecting morpheme *'s'*. The system would then decompose the unknown word into *Arbeit ('job, work')* which is still correct, and *Samt ('velvet')*, which is obviously not the expected second constituent (which has to be *Amt ('office, department, center')* because the *'s'* is not interpreted as a morpheme but as the first letter of the second constituent.[2] For several reasons, the correct encoding of the connecting morphemes (*Fugen-code*) is not as trivial as it might appear. First, there are various types of these elements: zero morpheme: *Umwelt* → *Umwelt-bewegung*; addition of a form of the inflectional paradigm of the word, e.g. the plural ending: *Diskette* → *Diskette+n-laufwerk*; addition of a letter which is not in the inflectional paradigm: *Installation* → *Installation+s-programm*; deletion of the ending: *Schule* → *Schul-hof*; deletion of the ending and addition of another letter: *Weihnachten* → *Weihnacht+s-konzert*.

There are quite a lot of words, however, which can take more than one type of connecting morpheme. In some cases, it is only a question of usage, depending on the head noun, in which form the word appears; in other cases, the type of juncture morpheme has significance in meaning distinction. The noun *Geschichte ('story/history')* is an example for such a case (cf. [Fleischer 1982]):

Geschicht+s–buch	–	'history book'
Geschichte+n–buch	–	'story book'

This fact which can help disambiguation has to be represented in the lexicon as a transfer constraint for compound nouns. The type of juncture element is not predictable from other formal aspects of the noun, e.g. from gender, declension code, etc. There are certain regularities, but they are not consistent enough to allow for an automatic encoding. It is just as little possible to derive the connecting elements completely from existing machine readable dictionaries (MRD); as a prerequisite, all words would have to appear in an MRD in all their possible forms as modifying elements of compound words.

The codes which we assigned to the connecting elements relate only to the form of the morpheme. As far as the implementation is concerned, the formal identity of some connecting elements and inflectional morphemes on the one hand is used to simplify the segmentation algorithm, and, on the other hand, the difference between connecting elements which are in the inflectional paradigm and those which are not is used to make predictions on the possible position of a constituent in a compound word.

2.3 Possible Positions of Compound Constituents

It is possible to draw certain conclusions from the type of connecting element on the possible position of a constituent in a compound word. Depending on whether the juncture morpheme has the same form as a form of the inflectional paradigm of the word or not, or whether the ending of the base form of the word is deleted or not, the word with its juncture can be positioned as a modifying constituent in the beginning or in the middle of the compound, or as the modified constituent (the head) at the end, or in any combination of the mentioned positions. The following examples will make the idea clearer.

□ Words with zero juncture can be at any position in a compound word:
Import–beschränkung ('import restriction')
Fisch–import ('fish import')
Fisch–import–beschränkung

□ Words of which the connecting element is in the inflectional paradigm can also be at any position in a compound word:
Parlament+s–debatte ('parliamentary debate')
(der Sitz des) Bundes–parlament+s
('(the seat of the) federal parliament')

□ Words of which the ending is deleted can only be in front or middle position: Schul–hof ('schoolyard'), *Musik–schul, but –schule ('music school')

□ Words of which the connecting element is not in the inflectional paradigm can only be in front or middle position:
Information+s–material ('inform. material')
*Studenten–information+s, but –information
('information for students')

2.4 The Segmentation Procedure of COMPGE in LMT-GE

The general framework for our research work and implementation is the machine translation system **LMT** developed by Michael McCord.[3] **LMT** is a lexicalistic, source-based transfer system. In this section, we concentrate on the performance of the PROLOG algorithm 'Compound Interpretation COMPGE' as a hook-up component to LMT-GE (German-English).

The segmentation and translation algorithm COMPGE is only called upon if an input word (with more than five letters) has not been found in the system's lexicon or in the on-line accessible MRD Collins German-English[4], i.e. when lookup and the regular

[2] More examples can be found in ([Luckhardt/Zimmermann 1991], 116f).

[3] LMT and related projects are described in detail in ([McCord 1989]; [Rimon et al. 1991]; [Schwall 1991]).

[4] For further information, cf. ([Neff/McCord 1990]).

morphological analysis fail. The segmentation is then carried out from left to right, beginning after the third letter. The decomposition process continues until the first word is found in the lexicon; the dictionary entry contains, among other data, information about the connecting element (*Fugen-code*). The algorithm then takes the complete dictionary entry with source and target word and all information contained in it, stores the word and continues by looking up the rest as a whole. If an entry is found, it is stored as well, together with the relevant morphological, syntactic, and semantic information. If there is, on the other hand, no entry for the remainder as a whole, the segmentation is carried on letter for letter, the same way as for the first constituent until an analysis for an existing entry is derived.

When all constituents are found, the words are stored, and segmentation is started again in order to allow, in ambiguous cases, for more than one possible segmentation. Let us look at the word *Messerattentat*. The result of the first decomposition would be *Messe-ratten-tat ('mass-rat-action')*, in accordance with the *Fugen-codes* of the segments; the second result would be *Messer-attentat ('knife-attack')*, also in accordance with the *Fugen-codes*. The system which then has to choose between the two possibilities would take the second result following the general strategy that compounds with two nominal constituents are much more frequent than those with three elements, those with three more frequent than those with four, etc. (cf. [Jeziorski 1982], [Müller 1977]). When segmentation is finished, the algorithm begins with the semantic interpretation of the compound before starting transfer.

2.5 Syntactic and Semantic Implications

Since, in non lexicalized compounds, the compound is generally a member of the syntactic and semantic class to which its head word belongs, this information can be passed on to the whole compound. As mentioned earlier, the entry for each constituent of the compound is extracted from the lexicon. Then the relevant morphological, syntactic and semantic information of the last constituent, the head noun, is attributed to the compound word as a whole. The following example *Umweltbewegung* illustrates the procedure: Whereas *Umwelt* has the semantic type **physical**[5], *Bewegung* gets the type **abstract**. Consequently, the compound word is attributed the semantic type **abstract**, too. This passing on of semantic information[6] can be used, for instance, for target lexeme selection using semantic constraints or for anaphora resolution.

3 Transfer and Generation

Transfer in LMT is divided into two parts: the compositional transfer which is part of the shell, and the language pair dependent restructuring transfer. The translation of compound words is done during compositional transfer.

In order to translate German compounds correctly into English, contrastive research studies had to be carried out on compounding phenomena. We first set up a typology of German and English **morphological** correspondences of compound constructions. Analysis was first done on the basis of 17,400 nominal compounds extracted from the MRD Collins German-English. In a second phase, in order to compensate for the fact that there are also lexicalized, nonproductive compound types in the dictionary, monolingual corpus based research was carried out (cf. 3.3).

3.1 Feature Transfer

Morphological and syntactic information on the source head word is passed on to the corresponding target word. If there is a specific feature of the target word coded in the transfer part of the lexicon which contradicts a source feature, the last one is overwritten by the target feature. If for instance the target word only occurs in the **singular**, but the source head word of the compound has the feature **plural**, the target word feature is preferred over the source word feature, and the compound will appear in the singular. e.g. the plural word *Industrieinformationen* becomes a singular in English — *industry information* — because of the transfer lexicon part
`< t(information)/sg.`
Other information that goes with the target head word entry such as information on subcategorization is passed on to the target compound construction as well.[7]

3.2 Analysis of the Compounds of a Bilingual Dictionary

The aim of our contrastive study was to find out correspondences between morphological types of German and English compound nouns. Therefore, a classification was set up where six types of German nominal compounds were contrasted with twelve types of English nominal compound constructions. These types contained information on the POS of the compound constituents, i.e. on the internal structure of the compounds in both languages.

After encoding 17,400 German compounds with their English correspondences according to these types, an evaluation was made which led to the following results: The *noun-noun* construction is the most frequent type in German as well as in English. What is even more important for the translation strategy is the fact that 54.4% of the German *noun-noun* compounds are translated into the same English construction type, i.e. into *noun-noun*-compounds as

[5] On the semantic type hierarchy used in LMT-GE, cf. [Breidt 1991].

[6] Since we intend to treat only non lexicalized compounds this way, a false semantic analysis — as it might occur in trying to translate the word *Frauenzimmer* (not 'women's room', but rather an archaic/derogatory term for 'woman') this way — is not very probable, given the fact that these kinds of words can be found in the LMT-lexicon or in on-line accessible dictionaries.

[7] In certain cases, a slot of the frame is filled by the modifier of the headnoun of a compound. Nevertheless, this is not always the case; therefore, we prefer passing on the subcategorization frame (cf. [Fanselow 1981]).

well. They are followed by the ***adjective-noun***-type (17.2%) and the ***noun-of-noun***-type (14.3%). Considering all German nominal compounds and not only ***noun-noun***-compounds, 44.4% of them were translated into the English ***noun-noun***-type.[8]

These are the data which formed the basis for our first translation strategy, namely to translate German nominal compounds per default into English ***noun-noun***-constructions. Since about 50% would then not be translated correctly, i.e. not according to language usage, this first approach has been augmented by corpus based techniques which are currently at an experimental level.

3.3 Corpus Based Techniques

3.3.1 Selecting the Target Construction

Recognizing that selecting the preferred target construction for a certain compound is in part an arbitrary decision of each language, it seems suitable to look for the information in a target language corpus. The idea is that when the English compound we should generate does not appear in the system's lexicon we will try to match it against the corpus and select a preferred construction according to the information found[9]. It should be noted at this point that in many cases there are several legitimate constructions that may be selected. However, the system cannot always distinguish these cases from cases where there is only one legitimate choice in the specific context. Therefore, it is always necessary to make a selection, and our strategy is to prefer the construction that seems most probable for being a legitimate choice. This strategy has also a stylistic advantage, as it prefers the more commonly used constructions.

The most simple and accurate method to start with is to search the corpus for explicit examples of the complete compound and prefer that construction which is most frequent. For instance, the German compound ***'Oppositionsgruppe'*** may in principle be translated (according to the findings described in the previous section) to either ***'opposition group'***, ***'group of opposition'***, ***'oppositional group'*** or ***'opposition's group'***. Consulting a corpus of 40 million words of The Washington Post articles enables us to prefer the first (***'noun-noun'***) option as it occurs 89 times in the corpus, while the second option occurs only 3 times and the other options do not occur at all. On the other hand, in translating the compound ***'Parlamentsdebatte'*** the statistics prefer the construction ***'parliamentary debate'*** (23 occurrences), where the modifier appears in its adjectival form. In this case, the ***'noun-noun'*** form, ***'parliament debate'***, does not occur in the corpus, and the form ***'debate in parliament'*** occurs 3 times.

In the cases mentioned above, the corpus provides enough examples of the exact compound we are looking for. The only generalization that was used is to take into account the morphological inflections of the words (e.g. counting also occurrences of ***'parliamentary debates'***, with the plural form of ***'debate'***). However, many compounds are too rare and do not occur a significant number of times in the corpus. In these cases it is necessary to use various generalizations over the constituents of the compound in order to obtain some relevant information. A suitable solution is to generalize over the part of speech of some of the words of the compound. For example, the compound ***'Umweltbewegung'***, may be translated (among other options) to ***'ecology movement'*** or ***'ecological movement'***. This compound occurs only once in The Washington Post corpus, in the form ***'ecological movement'***, but this is not significant enough to make a selection. In order to obtain more information we look for compounds in which either ***'ecology'*** or ***'ecological'*** serves as a prenominal modifier, with no restriction on the specific word which serves as the head noun. This information was searched for in the first 100,000 sentences of the Hansard corpus of the proceedings of the Canadian parliament, which was tagged with part of speech using a stochastic tagger [Merialdo 1991]. In these sentences, the form ***'ecological ⟨noun⟩'*** was observed 11 times while the form ***'ecology ⟨noun⟩'*** only once. Using these statistics we regard the adjectival form ***'ecological'*** as the default form whenever the two alternatives are encountered and there are not enough examples of the complete compound. For instance, this default will be used also when translating ***'Umweltprobleme'*** to ***'ecological problems'*** or ***'Umweltreserven'*** to ***'ecological reserve'*** (and not inappropriately to ***'ecology problems/reserve'***). The use of such defaults enables us to increase the coverage of the statistical method and treat infrequent compounds of the target language.

Another important purpose for using default constructions for single words is to save storage space. Without defaults, we would have to store in our statistical data base the most frequent construction for every specific compound which occurs in the training corpus a significant number of times. This might require too much space when training the system on the very large corpora which are necessary to get high coverage and precision of the method. On the other hand, if we store the default constructions for single words, then we should store specific compounds, i.e. combinations of words, only when the preferred construction for these combinations conflicts with the defaults for single words.

This leads to the following implementation scheme: During the training phase, the (tagged) corpus will be processed twice. In the first pass, the default constructions for single words will be collected. In the second pass, all the specific compounds will be identified, but only those which conflict with the default constructions will be stored in an exception list. When translating a new German compound (during the actual translation phase), the exception list will first be consulted to check whether one of its items matches one of the possible alternatives for translation. Only if there is no relevant item, the defaults for the single constituents will be used.

3.3.2 Selecting the Target Lexemes

We relate to the problem of selecting the appropriate target words for the constituents of the compound

[8] The contrastive studies and their results are described in detail in [Rackow 1992].

[9] This approach is applicable for any natural language generation task, hence the relevance of this section is not restricted to the application of machine translation.

as a special case of the problem of target word selection in machine translation (which itself is a variant of lexical disambiguation). As such, these ambiguities will be treated by the general method described in [Dagan et al. 1991], which uses statistical data on lexical cooccurrence within specific syntactic relations in a target language corpus.

Consider the following example given for illustration. The German compound *'Reformprozeß' ('reform process')* has in principle 9 possible translations. There are three possible English constructions, *'noun-noun, noun of noun, noun's-noun*, and three possible translations for the word *'Prozeß', 'process', 'case'* and *'trial'*. Out of these 9 alternatives, the compound *'reform process'* occurs 5 times in the first half of The Washington Post corpus, while all the other alternatives (*'process of reform', 'case of reform', 'reform case'* etc.) never occur. Using these statistics, the algorithm described in [Dagan et al. 1991] selects *'reform process'* as the preferred translation. It should be noted that the information which is used for lexical disambiguation may come from either within the compound, as in this example, or from the surrounding context, such as using the verb which interacts with the compound.

4 Conclusions

This paper demonstrates that the translation of noun compounds is a difficult task. Having German as the source language adds the problem of segmenting the compound into its constituents, a problem which does not exist in many other languages. The solution for these problems seems to require various levels of information, involving morphological, syntactic, semantic and stylistic criteria.

Though these levels are general for every natural language processing task, we have shown how a detailed analysis of the specific linguistic phenomena can lead to an efficient hybrid architecture which uses the partial information available computationally. This architecture combines formal syntactic and morphological rules, wherever they can be specified accurately, with empirical data which reflects some of the semantic and stylistic considerations. In this sense, this paper promotes the integration of the sometimes diverging streams, namely the use of symbolic, manually stipulated linguistic rules versus the use of statistical data which is extracted automatically from corpora. In our view, these two disciplines complement each other and are both essential to achieve high performance in practical natural language processing systems.

Acknowledgements: We would like to thank Eran Amir from Haifa, Peter Brown from Yorktown and Mark Beers and Myriam Welschbillig from Heidelberg for their help and comments.

References

[1] S. Ananiadou and J. McNaught. Treatment of Compounds in a Transfer-based Machine Translation System. In *Proc. of the 3rd Int. Conf. on Theoretical and Methodological Issues in MT of NL*, Univ. of Texas, Austin, 1990.

[2] H. U. Block. *Maschinelle Übersetzung komplexer französischer Nominalsyntagmen ins Deutsche.* Niemeyer, Tübingen, 1984.

[3] E. Breidt. *Die Behandlung von mehrdeutigen Verben in der maschinellen Übersetzung.* IBM IWBS Report 158, Stuttgart/Heidelberg, 1991.

[4] I. Dagan, A. Itai, and U. Schwall. Two languages are more informative than one. In *Proc. of the 29th Meeting of the ACL*, pages 130-137, Berkeley, 1991.

[5] P. Downing. On the creation and use of English compound nouns. *Language*, (53):810-842, 1977.

[6] G. Fanselow. *Zur Syntax und Semantik der Nominalkomposition.* Niemeyer, Tübingen, 1981.

[7] W. Fleischer. *Wortbildung der deutschen Gegenwartssprache.* Niemeyer, Tübingen, 1982.

[8] J. Jeziorski. Strukturmodelle der deutschen Nominalkomposita vom Typ 'Substantiv + Substantiv'. *Wirkendes Wort*, (4):235-238, 1982.

[9] H.-D. Luckhardt and H. H. Zimmermann. *Computergestützte und Maschinelle Übersetzung.* Saarbrücken, 1991.

[10] M. C. McCord. The slot grammar system. In J. Wedekind and Ch. Rohrer, editors, *Unification in Grammar*, MIT Press. to appear.

[11] M.C. McCord. A New Version of the Machine Translation System LMT. *Literary and Linguistic Computing*, (4):218-229, 1989.

[12] B. Merialdo. Tagging text with a probabilistic model. *ICASSP*, 1991.

[13] B. S. Müller. Einige statistische Angaben über zusammengesetzte Substantive im Deutschen. *Germanist. Linguistik*, (1/2):171-198, 1977.

[14] M. Neff and M. McCord. Acquiring lexical data from machine-readable dictionary resources for machine translation. In *Proc. of the 3rd Int. Conf. on Theoretical and Methodological Issues in MT of NL*, pages 87-92, Univ. of Texas, Austin, 1990.

[15] H. Ortner and L. Ortner. *Zur Theorie und Praxis der Kompositaforschung.* Narr, Tübingen, 1984.

[16] U. Rackow. *On the Treatment of Compounds in Machine Translation. A Study.* IBM IWBS Technical Report, Heidelberg, 1992. To appear.

[17] M. Rimon, M. McCord, U. Schwall, and P. Martínez. Advances in Machine Translation Research in IBM. In *Proceedings of the MT Summit III*, pages 11-18, Washington, 1991.

[18] U. Schwall. *LMT-Machine Translation Demonstration.* IBM IWBS Report 177, Stuttgart/Heidelberg, 1991.

[19] Eiichiro Sumita and Hitoshi Iida. Eperiments and Prospects of Example-based Machine Translation. In *Proc. of the 29th Meeting of the ACL*, pages 185-192, Berkeley, 1991.