Extracting Phrasal Terms using Bitext

Jörg Tiedemann
Department of Linguistics
Uppsala University
Box 527
S-75120 Uppsala
Sweden
Tel +46 (0)18-471 7007
Fax +46 (0)18-471 1416
joerg@stp.ling.uu.se

Abstract

This paper focuses on the improvement of automatically generated phrase lists by applying word alignment approaches to parallel bitext. Such phrase lists, in terms of multi-word collocations, serve several tasks such as the compilation of terminology databases or translation database in the multilingual case. Our investigations are based on the assumption that word alignment favors well-formed phrase structures rather than irregular text segments. If this is the case, word alignment will filter out irregular structures from automatically generated phrase lists. As a result, an improved phrase list may be compiled based on the bilingual lexicon that could be extracted. Furthermore, word alignment approaches can be used to identify additional multi-word units by comparing corresponding multiword units from the bitext. Our investigations will be focused on a Swedish/English text collection that has been aligned with the Uppsala Word Aligner (UWA).

1 Introduction

Domain specific terminology is one of the most important language resources in document production. Consistency of terms and their usage has to be assured especially in public documents such as political and technical texts. Authors of such documents need to be supported by efficient tools and comprehensive information about terms and their appropriate usage. Furthermore, an increasing amount of documents has to be translated into several languages. However, translation is often carried out by external partners, which is time and cost-intensive. In addition, external translation partners are usually not familiar with the terminology used for specific purposes. Therefore, the need of term databases becomes even more important for translation processes than for document production.

In recent years much effort was devoted to automate the process of compiling terminology by analyzing document collections. Valuable information, monolingual and multi-lingual, is included in huge amounts in previously written documents. Such information can be used for building term databases that support document production, human and (semi-) automatic translation.

However, the task of identifying and extracting terminology information is not trivial. Usually, single-word approaches produce reliable results with little effort. Difficulties arise with phrasal terms in both monolingual and multi-lingual approaches. The problem can be defined as the problem of identifying appropriate multi-word units (MWUs) that are part of the terminology. However, the definition of phrasal terms is not straightforward and may change for different applications. Phrasal terms for translation purposes may include common noun phrases and other types of grammatical phrases whereas terminology databases often focus on proper names and non-compositional compounds.

This paper will focus on the bilingual case, i.e. the extraction of phrasal terms that can be linked to correspondences in another language. Such term collections include non-compositional compounds, multiword names and labels, idiomatic expressions, and so on.

Related work on the acquisition of phrasal terminology can be found in (Dagan and Church, 1994), in which the authors propose a semi-automatic tool for the identification and translation of technical terminology. Another investigation on bilingual terminology that includes a phrase-based model is described in (van der Eijk, 1993). In (Smadja et al., 1996), an approach is presented that combines monolingual phrase generation with statistical word alignment. Another work that focuses on the identification of non-compositional compounds is presented in (Melamed, 1997).

Our investigations are based on word alignment approaches by means of the Uppsala Word Aligner (UWA)[1] (Tiedemann, 1998; Tiedemann, 1999b). Experimental results will be taken from the Swedish/English portion of the Scania corpus,

[1]UWA was developed within the co-operative project on parallel corpora, PLUG (Sågvall Hein, 1999). PLUG was funded by "The Swedish Council for Research in the Humanities and Social Sciences" HSFR and the "Swedish National Board for Industrial and Technical Development" NUTEK.

a multi-lingual collection of technical text provided by the Swedish truck and bus manufacturer Scania. This corpus is currently used for the definition of a controlled language for Scania to be used in the document production in the future (Almqvist and Sågvall Hein, 1996).

2 The basic assumption

Terminology extraction is commonly based on monolingual text collections. Significant phrases, which are part of the terminology, are particularly hard to detect due to their indistinct characterization. However, an increasing amount of textual data is available that is translated into several languages. With the assumption that phrasal terms are translated consistently into other languages we suppose that bilingual word alignment can be used to improve the quality of automatic term extraction from text for both languages. In this paper we will focus on the application of the Uppsala Word Aligner to a Swedish/English bitext in order to investigate phrase extractions at different levels of the alignment process.

3 Handling phrases in UWA

Word alignment is one of the tasks in corpus linguistics where phrasal structures have to be handled. Within the co-operative project on parallel corpora PLUG we experienced the complexity of word alignment approaches and their evaluation when phrasal structures are involved. The Uppsala Word Aligner (UWA) handles the bilingual alignment of phrasal structures in the following way:

1. Generate phrases for both languages from the text under consideration.
2. Annotate phrases in the text.
3. Consider phrases as single tokens for co-occurrence measures.
4. Align multi-word units if they fulfill certain conditions.

3.1 Phrase generation

The automatic generation of phrases in UWA is based on word association scores. Similarly to (Smadja, 1993), an iterative process is applied in which the size of word N-grams is increased step by step. We apply mutual information scores in order to determine the significance of the current unit, i.e. the significance between the current (N-1)-gram X and its direct successor Y:

$$MI = log_2 \frac{prob(X_{N-1}|Y)}{prob(X_{N-1})prob(Y)} \quad (1)$$

Furthermore, we use classified stop word lists, as proposed in (Merkel and Andersson, 2000) in order to exclude certain functional words from certain positions in valid phrases and to define break points for phrasal structures. In this way, certain words can be excluded from the beginning of phrases, the end of phrases, from within phrases, or from any phrase. Another parameter is the over-all frequency of the phrase within the complete text. Currently, only contiguous bi-grams and tri-grams are considered in UWA for efficiency reasons.

Thanks to the combination of association scores, classified stop word lists, and frequency thresholds this technique produces valuable phrase collections within reasonable processing time. However, the iterative nature produces many inclusions, incomplete units and overlapping units. Consider the following example phrases that have been extracted from the Scania material:

```
ABS warning
ABS warning lamp
axle raise
axle raised
front axle
tag axle
```

Here, 'ABS warning' represents a typical inclusion and might be incomplete as a phrase because of the missing word 'lamp'. The last four examples show typical overlapping phrases that have been extracted. The word 'axle' is common for all the four MWUs and overlaps in contexts such as 'the front axle raised'.

3.2 Annotating phrases

The next step in UWA alignment processes is an annotation step. Here, UWA uses phrase lists and tries to identify and annotate them in the text. In our case, automatically generated phrases as described above are applied for this purpose. The UWA annotation module applies a simple left-to-right segmentation and annotates the longest (counted in words) valid unit that can be found in the text and continues the annotation process with the subsequent word of the last annotation. With this heuristic, overlapping phrases and inclusions are excluded from the phrase list. Consider the following examples from the Scania corpus that have been annotated using the phrase list from the example above.

Without **front axle** raised
... and the **ABS warning lamp** lights.

Annotation is done for both languages. New phrase lists can be extracted from the annotated text in which inclusions and overlapping phrases are excluded.

3.3 Aligning phrases

Up till now, all processing could have been done for monolingual texts as well. Now, we will take advantage of the parallel character of the corpus. We assume that phrasal expressions tend to be translated consistently into corresponding expressions in other languages. According to this assumption, we suppose that phrasal terms will be aligned more confidently than insignificant word sequences such that the quality of the phrase list will be improved.

UWA takes bilingual text (bitext) that has been sentence aligned previously as its input. Now, the system calculates association scores for possible combinations of tokens (including annotated phrases) from sentence alignments in order to collect candidates of translation equivalents. Furthermore, UWA uses additional extraction approaches such as investigations on string similarity (Tiedemann, 1999a; Tiedemann, 1999b). This technique can be applied to multi-word units as well and in this way, phrasal cognates can be identified. Consider the following Swedish/English examples:

```
varningslampor    se  exempel
warning lamps     see example

   retardens oljefilter
   retarder  oil filter
```

Using all these candidates, rated by their association score, the actual word alignment begins, starting with the first sentence pair. First, UWA generates all possible word combinations for both languages[2] and combines pairs of such combinations from the two sentences within a certain link window. The maximum length of word sequences can be adjusted by appropriate parameters. Currently, the maximum length is set to three. Secondly, the system searches each pair in the list of translation candidates. All pairs that could be found are stored in the link list sorted by their score. Finally, the system aligns all pairs starting with the pair with the highest score. In this way, the most confident pairs are linked first. All linked words are removed immediately from the sentences such that no word can be aligned twice.

The result of UWA alignments is a list of links and an extracted bilingual lexicon. Phrasal terms are included for both the source and the target language. These terms include phrases with a certain translation consistency. Furthermore, additional phrasal terms could be identified by additional alignment techniques such as approaches using string similarity measures.

[2] Only contiguous sequences are supported.

Scania 2000	*Swedish*	*English*
generated phrases	16,496	23,312
annotated phrases	12,188	17,246
aligned phrases	4,138	10,831

Table 1: Phrase extractions from the Scania corpus.

4 Evaluating phrase extraction results

Phrasal terms have been produced for two languages on three different levels as the result of the previously described alignment process:

- automatically generated phrases
- phrases that have been annotated
- phrases that have been aligned

Each step excludes insignificant phrases discarding over-generations. Furthermore, bilingual word alignment produces additional phrases such as cross-lingual phrasal cognates that have been overlooked in the monolingual phrase generation.

Evaluating quality and completeness of such lists is by far not trivial. As mentioned earlier, the purpose of the application of such lists is an important evaluation criterion.

For evaluation purposes, we aligned the complete Swedish/English portion of the Scania corpus with about 2.7 million words using UWA with the techniques described above. Table 1 summarizes the number of extracted phrases at each of the three levels: automatically generated phrases, annotated phrases, and bilingually aligned phrases.

A qualitative evaluation of these phrases can be made in different ways. One method is to manually check a representative part of each extracted phrase list. Another possibility is to define valid phrase patterns in terms of part-of-speech sequences and compare these patterns with the resulting phrases. The latter requires tagging of extracted phrases. We decided to apply the second approach to our evaluations. For the tagging task we apply the Uppsala Chart Parser (UCP) including a corpus specific Swedish stem lexicon (Sågvall Hein, 1995). The results of this tagging process are lists of tag-sequences that are associated with the phrase lists that have been compiled.

The next step comprises the definition of tag-sequences for valid phrase constructions. This task is by far not trivial due to the huge variety of phrasal constructions that can be included in natural language texts. The main task is to identify patterns of the most common and most desirable phrases that are expected from the extraction process. Our investigations build on former studies on phrase structures within the Scania corpus

```
##############################################
# well-formed phrases
##############################################
(AV.*|NN.*)             # olika problem
(AL.*|AV.*|NN..[IX].*) # ett tomt system
(NN...[GX].*|NN.*)      # bussens sidolucka
(AV.*|AV.*)             # helt rund
(VB.[PRM].*|AB.*|NN.*) # tryck ner gaspedalen
(VB.*|AB.*)             # sitta kvar
##############################################
# partial phrases
##############################################
(VB.I.*|NN.*)           # styra fordonet
(AV.*|NN.*|VB.[PR].*)  # nedre ringarna finns
##############################################
# text-specific:
##############################################
(PM.*|PM.*)             # Scania Holland
(NN.*|NN.*)             # ABS systemet
(TYP.*|NN.*)            # M1 minne
```

Figure 1: Examples of phrase patterns for the Swedish text from the Scania corpus.

and the Swedish Statement of Government Policy (Sågvall Hein et al., 1990; Wikholm et al., 1993). Additionally, we include some text-specific partial structures that might be of interest for applications like machine translation and terminology compilation; they contribute valuable valency information and semantic relations between domain-specific expressions. Furthermore, even multi-word-units that violate certain language specific rules for phrase construction might be included if they follow a frequent pattern in the text. Consider the examples of some patterns of Swedish phrase structures from the Scania corpus which are illustrated in figure 1.

The first part of each tag specifies a part of speech and the following letters specify morpho-syntactic features. The definition is straightforward, taken from parsing results of the UCP, with *AV* denoting adjectives, *NN* denoting nouns, *VB* denoting verbs, *AV* denoting adverbs, *PM* denoting proper nouns, and *TYP* denoting labels in the example in figure 1. Tags in the sequences are separated by "|" characters. The syntax of tag patterns follows the rules for regular expressions such as used in the programming language Perl. In most cases, phrase patterns simply describe sequences of part-of-speech-tags. The values of additional features are used in certain cases in order to exclude invalid structures that would match a more general pattern. An example is pattern two, which requires indefinite forms of nouns at the third position in the phrase (feature 3 is set to "[*IX*]"). Pattern three requires nouns in genitive case ("[*GX*]") at position one. The first tag in pattern five matches finite verbs only, which are inflected in present tense ("P"), past tense ("R"), or represent imperative forms ("M").

Swedish phrases	tagged	correct	invalid
generated	15,317	81.17%	18.83%
annotated	11,442	85.14%	14.86%
aligned	3,620	88.59%	11.41%

Table 2: Evaluating phrase extractions using phrase pattern.

The partial phrase patterns in figure 1 accept verbal phrases without the modal verb at the initial position. These phrases might be used in order to determine the valency of verbs.

Text-specific phrase patterns refer to pattern that recur frequently in the Scania corpus although their structure is invalid in Swedish. These pattern pop up as competence errors in the document production. The example in figure 1 represent examples where the authors violated the general rules for Swedish compounding. However, these structures might be interesting e.g. for grammar-checking and the extraction of translation equivalents. Therefore, they have been included in our example as well.

The list of phrase patterns was applied in order to evaluate the phrase lists that were compiled by the system. Each phrase list was tagged as described above including alternative tags in case of ambiguity. Each phrase that could be matched with one of the phrase patterns has been marked as *correct*, each phrase that could not be tagged completely was marked as *unknown*, and all others were marked as *invalid*. Table 2 illustrates the results for each of the three extraction levels.

The numbers in table 2 describe a clear quality differences of phrase lists that have been extracted at different levels by the UWA system. The improvements might be even more apparent if inclusions would be evaluated in the evaluation. Valid subphrases that should not be included in such phrase lists occur rather frequently in the generated phrase list. Such inclusions are filtered out during the alignment process because UWA does not allow ambiguous and overlapping alignments.

The numbers in table 2 refer to completely tagged phrases only. Incompletely tagged phrases are hard to judge although they might tend to be invalid. However, even if all phrases that have been marked with "unknown" would have been counted as invalid, an improvement from 75.37% (generated) to 79.93% (annotated) correct phrases could have been measured. The list of aligned phrases includes a rather large portion of incompletely analyzed phrases (12.52%) such that the result would be decreased to 77.5% correct phrases although the portion of phrases with the "invalid"-marker was de-

Swedish phrases	correct	invalid
generated but not annotated	69.45%	**30.55%**
annotated but not aligned	83.81%	**16.19%**
aligned but not annotated	**86.58%**	13.42%
annotated and aligned	**89.13%**	10.87%

Table 3: Comparing phrase extractions at different stages.

creased drastically (9.98% compared to 17.48% from the generated phrase list).

The precision estimations in table 2 refer to the quality of each phrase list only. Another important parameter, usually referred to as "recall", is the portion of relevant phrases that could be obtained. However, the total number of relevant phrases in a document is hard to estimate. An indication on how much the recall value differs between the three levels can be seen in the the numbers of phrases that could be extracted. Word alignment is rather incomplete especially in case of phrasal structures. This fact can be seen in the number of remaining phrases from the alignment phase. The amount of phrases has been reduced drastically compared to the steps before. However, high precision values might be more important for tasks such as terminology compilation.

Another way of judging the reductions between the three phases comprises the analysis of phrases that have been excluded in each step. Table 3 illustrates the quality of intersections between the three sets of phrases.

Overlapping phrases and inclusions are excluded from the phrase list within the annotation phase. A rather large portion of them has been marked as invalid such that the quality of the remaining phrase list has been improved as shown in table 2 above. This also approves that the annotation process tends to divide the text into valid segments.

The alignment process, which applies the annotated text, fails to align a large amount of valid phrases. However, phrases that are actually aligned represent mostly valid phrases (according to the evaluation method) as can be seen in the last two rows in table 3. This statement is true for the 989 newly obtained phrases (about 24% of all alignments) that have been aligned but not annotated in the previous step.

UWA supports iterative processing as described in (Tiedemann, 1999b). However, all results above are taken from one iteration only. The word alignment process was carried out in three iterations for the complete Scania corpus. UWA removes all units that have been aligned from the text. For our investigations, the system was adjusted to generate new phrase lists after each iteration from the remaining text. The following values summarize the quality of the complete phrase lists after three iterations:

		tagged	correct
generated	21,991	20,519	76.94%
aligned	4,564	4,056	88.56%

Phrases that have been generated but not aligned were invalid in 25.37% of the cases whereas generated phrases that have been aligned were correct in 88.67% of the cases. These values approve a clear improvement of the quality of the resulting phrase list by applying word alignment. However, the number of aligned phrases has been reduced drastically.

5 Conclusion and Outlook

The basic assumption made in section 2 was substantiated for our investigations on Swedish/English bitext. The quality of the phrase lists could be improved significantly by applying automatic word alignment approaches. However, the number of extracted phrases is reduced drastically such that the resulting phrase collection is less complete than the originally generated phrase list. On the other hand, high precision values seem to be more important for the purpose of terminology compilation than completeness. Word alignment can support the process of filtering out invalid phrase constructions as shown in this paper.

Another advantage is the alignment of such phrases to their cross-lingual equivalents. Such databases are essential for machine translation and translation support.

Our investigations were based on Swedish phrases. However, the extraction of phrasal terms becomes even more important for languages like English which tend to express terms with non-compositional compounds. Here, word alignment produces many of such constructions especially in combination with languages like Swedish that require concatenation for creating compounds. Lots of links between compositional compounds and their non-compositional equivalents can be established by word alignment. Investigations on English phrases will be carried out in future.

References

Ingrid Almqvist and Anna Sågvall Hein. 1996. Defining ScaniaSwedish - a Controlled Language for Truck Maintenance. In *Proceedings of the First International Workshop on Controlled Language Applications*, KU Leuven, Belgium.

Ido Dagan and Kenneth W. Church. 1994. Termight: Identifying and Translating Technical Terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, Stuttgart/Germany. Association for Computational Linguistics.

I. Dan Melamed. 1997. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP97)*, Providence/RI.

Magnus Merkel and Mikael Andersson. 2000. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In *Proceedings from RIAO, Paris, to appear*, Paris/France.

Anna Sågvall Hein, Annette Östling, and Eva Wikholm. 1990. Phrases in the Core Vocabulary. Technical report, Department of Linguistics, University of Uppsala.

Anna Sågvall Hein. 1995. Swedish morphology in Sve.UCP. Technical report, Department of Linguistics, University of Uppsala.

Anna Sågvall Hein. 1999. The PLUG Project: Parallel corpora in Linköping, uppsala, göteborg: Aims and achievements. In *Proceedings of the Symposium on Parallel Corpora, to appear*, Department of Linguistics, Uppsala University, Sweden.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translation Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics, 22(1).*

Frank Smadja. 1993. *Retrieving Collocations from Text: XTRACT.* Computational Linguistics.

Jörg Tiedemann. 1998. Extraction of translation equivalents from parallel corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALI98*, Center for Sprogteknologi and Department of General and Applied Linguistics, University of Copenhagen.

Jörg Tiedemann. 1999a. Automatic Construction of Weighted String Similarity Measures. In *Proceedings of the Joint SIGDAT Conference on Emperical Methods in Natural Language Processing and Very Large Corpora*, Univrsity of Maryland, College Park/MD.

Jörg Tiedemann. 1999b. Word Alignment - Step by Step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics NODALIDA99, to appear*, University of Trondheim, Norway.

Pim van der Eijk. 1993. Automating the Acquisition of Bilingual Terminology. In *Proceedings of the 6th Conference of the European Chapter of the ACL*, Utrecht/The Netherlands. Association for Computational Linguistics.

Eva Wikholm, Annette Östling, and Ingrid Maier. 1993. A Multilingual Dictionary of Functional Core Phrases with Prepositions. Technical report, Department of Linguistics, University of Uppsala.

A Example alignments of phrasal terms

11-motor	11-series engine
ABS informationslampa	ABS information lamp
ABS-reglerventil	ABS control valve
ABS-varningslampa	ABS warning lamp
ABS-varningslampan	the ABS warning lamp
Anslut luftslangen	Connect the air hose
Användarinstruktion	User instructions
Arbetsbeskrivning	Job description
Arbetsbeskrivning	Work Description
Arbetstryck	Operating pressure
Arbetstryck	Working pressure
avgasbroms	exhaust brake
avgasledning	exhaust pipe
avgasrör kpl	exhaust pipe complete
avgasläckor	exhaust leakage
avgassystem	exhaust systems
avgasutsläppen	exhaust emissions
Batterifrånskiljare	Battery disconnector
Batterifrånskiljare	Battery isolator
Batterifrånskiljare	Battery master switch
Bosch P4 ABS	Bosch P4 ABS
Bromsljuskontakt	Brake lamp switch
Bromsljuskontakt	Brake lights switch
blinkande punkt	a flashing dot
Låt motorn gå	Run the engine
Magnetskiva	Magnetic disc
Magnetskriva	Magnetic plate
Magnetspolar	Magnetic coils
Magnetstativ	Magnetic foot
Magnetventil	Solenoid valve
Manuell nivåreglering	Manual level control
Matningsspänningen	The power supply
Matningsspänningen	The supply voltage
Motorbromsprogram	Engine brake program
Motorolja SAE 10W-30	Engine oil SAE 10W-30
Motorolja SAE 10W-30	Motor oil SAE 10W-30
Märk upp cylinderhuvudena	Mark the cylinder heads
Sätt dit växellådan	Fit the gearbox
Sätt dit växelstången	Attach the gearshift
Sätt ihop kontaktstycket	Assemble the connector
Sätt tillbaka fläkten	Refit the fan
Sätt upp avdragare	Fit puller
Sätt upp navet	Mount the hub
Sätt upp navet	Secure the hub
Ta fram felkoder	Accessing the fault
Ta fram felkoder	Reading fault codes
Ta försiktigt	Carefully remove
Ta hjälp	Get help
Ta isär elanslutningarna	Unplug the electrical connections
Ta isär elkontakten	Disconnect the electrical connector
Ta isär elkontakten	Dismantle the connector
Ta loss kardanaxeln	Detach the propeller
Tryck lätt	Lightly press