Terminology Extraction and Term Ranking for Standardizing Term Banks

Magnus Merkel
Department of Computer and Information Science
Linköping University
Linköping, Sweden
magme@ida.liu.se

Jody Foo
Department of Computer and Information Science
Linköping University
Linköping, Sweden
jodfo@ida.liu.se

Abstract

This paper presents how word alignment techniques could be used for building standardized term banks. It is shown that time and effort could be saved by a relatively simple evaluation metric based on frequency data from term pairs, and source and target distributions inside the alignment results. The proposed Q-value metric is shown to outperform other tested metrics such as Dice's coefficient, and simple pair frequency.

1 Introduction

Quality assurance (QA) of products and services is standard procedure in most industrial areas today. In the area of document production and localization, quality assurance has been deemed to be both time-consuming and costly as most of the linguistic quality assurance has to be made manually. Of course, if used, spell and grammar checkers, and controlled language checkers for assuring that manuals are created using a special variety of Simplified English, will identify some errors, and sometimes also help to correct them. The major problem for technical writing and localization quality is to be found in inconsistent use of terminology. Or as Sue Ellen Wright puts it: "The primary source of rework is inconsistent terminology" (Wright 2006). Inconsistent terminology is perhaps most crucial in source language documentation (originals) as mistakes there will multiply by every translation. Lombard (2006) illustrates this phenomenon by an example where an American software development company may use a great variety of terms to refer to a closing or stopped application by inconsistently using terms such as *cancel, quit, close, end* and *stop* in the user interface and in the accompanying documentation. This is, as Lombard puts it, not a problem for the development team as they know what all these terms mean. The translators/localizers, however, will be tempted to translate every distinct source term choice to distinct target terms, thereby multiplying the inconsistency.

Poor quality in documentation could result not only in dissatisfied clients and users, but also in substantially increased costs for revisions, retranslations and delays. In addition, legal damages could make things worse, for example through lawsuits for serious factual mistakes in the source documentation or in translations. Capturing mistakes in documentation before they reach the users/readers is the only way to avoid the extra costs and inconvenience that poor quality will yield.

The obvious solution to the inconsistency dilemma can be found in a *standardized term bank*. Creating a term bank is very time consuming if it is done in the old way, i.e. by hand. During the last decade word alignment techniques have been used to create practically usable resources for translation activities, much faster than the manual way. However, as word alignment can never produce 100 per cent accurate term pairs, methods of how to filter out erroneous entries and efficiently revise the output from alignment systems need to be developed. Even if an alignment system is close to perfect, the data itself (the source and target texts) will contain errors, omissions and additions that will result in terminological entries that are unwanted in a standardized term bank. Perhaps most interesting, high quality alignments will produce a map of the

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 349–354

source and target texts that reveals how consistent, or inconsistent, term usage is in reality.

In this paper we will address the issue of how to create standardized term banks by using word alignment techniques. The focus lies on the ranking of the term entries produced by the alignment system and on the evaluation of a proposed metric.

2 Motivation

The focus of this paper is to explore how term candidate validation can be improved by using a good ranking metric. A good ranking metric correlates to the precision of a term candidate. This means that using a good term ranking metric makes it possible to select a set of term candidates which when processed will result in a higher number approved terms compared to selecting the set of term candidates to be processed by random or using a bad ranking metric.

3 Approach

The ranking metrics used here are based on data from a set of word aligned term candidates. This means that the presented method can be used on all term candidate sets, without regard to how the term candidates have been extracted or produced. A corpus-based approach relies on the existence of a corpus from which statistics can be calculated.

The set of term candidates used in this paper were extracted using an align-filter method. The extracted term candidates are then ranked using term pair frequency, Q-value and Dice's coefficient (see Metrics below). The ranking order produced by these metrics is then compared using accumulated precision.

The word alignment system used is the ITools suite, developed at Fodina Language Technology and Linköping University (Ahrenberg et al. 2003; Deléger et al. 2006; Nyström et al. 2006; Foo & Merkel 2007). The source and target texts used in the alignment case study consisted of around 35,000 sentence pairs (English-Swedish) from patent texts from the subject area *Animal care*.

3.1 The ITools suite

The ITools software suite three major applications interactive word alignment (ILink), automatic word alignment (ITrix), viewer for editing and browsing alignment data (IView).

The ITools suite also includes functions for sampling test and training data sets, automatic evaluation, statistical processing and conversion from XML to SQL database format.

The basic approach used for alignment in the ITools suite combines evidence from a variety of different sources by assigning each piece of evidence a score and then calculating a joint score for all of them (cf. Tiedemann 2003).

The ITools suite is supported by Connexor's Machinese Syntax parsers (Tapanainen & Järvinen 1997) which provide the grammatical information for English, Swedish and several other western European languages.

A typical word alignment process using the ITools suite consists of the following steps:

1. Morphological, syntactic and dependency analysis of source and target files
2. Statistical processing of source and target files
3. Sampling test and training data sets
4. Training, i.e. creating dynamic resources interactively (ILink)
5. Running automatic alignment (ITrix)
6. Conversion to SQL database
7. Verification, filtering and categorization of extracted term candidates (IView)

In step 1 the sentence aligned source and target text are parsed independently using the Machinese Syntax parsers for the source and target languages.

In step 2 statistical resources are created both for the word form level (inflected words) and lemma level (base forms). We use t-score and dice associations on co-occurrences between items in the bitext and thereby create a bilingual dictionary which is used as a static resource in the automatic alignment. Other statistical approaches can also be used, such as the Giza++ kit (Och & Ney 2005). In the third stage, a test set and a training set of aligned sentence pairs are randomly sampled. The size of these sets varies depending on the project and time available.

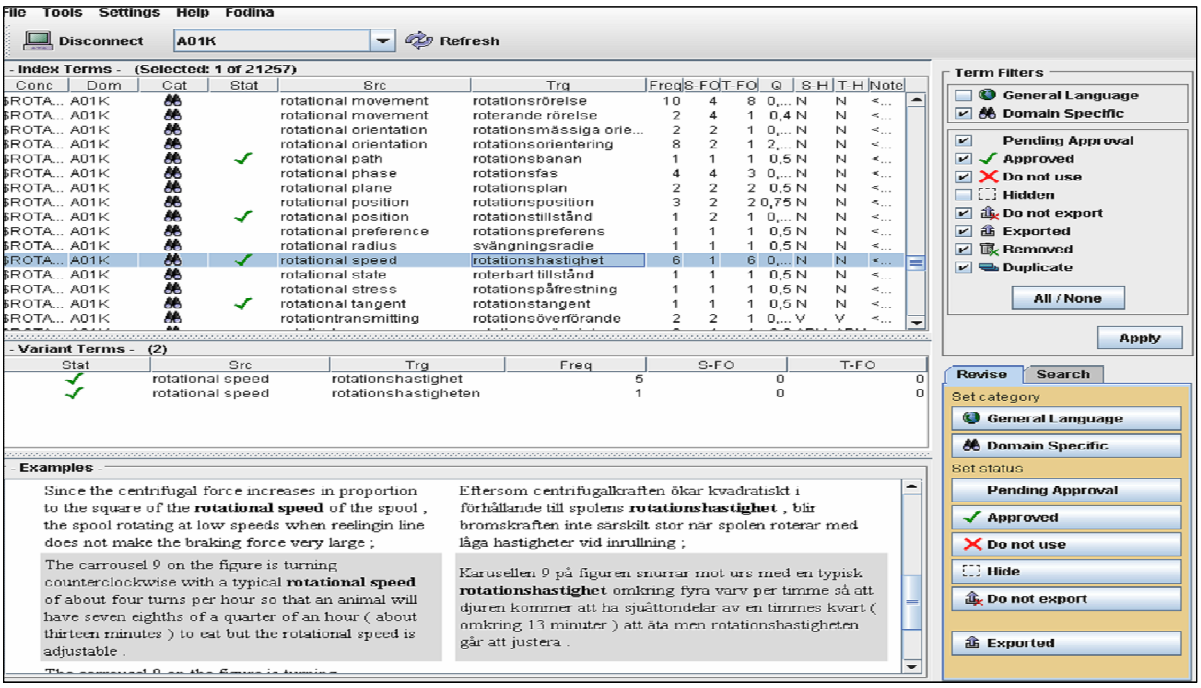

Figure 1. IView application. Used for filtering, categorization and revision.

In step 4, a training environment is set up in the interactive ILink tool where the training results in dynamic resources on four levels: 1) the word form level, 2) the base (lemma) level, 3) the parts-of-speech level, and 4) the syntactic function level.

In step 5 the automatic alignment is performed using ITrix, which results in thousands of pointers between the source and target texts containing the actual token links. These token links are then used to create an SQL database, keeping all grammatical information from the XML files as well as creating a structured term data base containing a concept level, index term level, term variant level and examples (see Figure 1).

However, to arrive at a usable term collection, the output from the word alignment needs to be verified. The IView application can be used during this last step (see Figure 1), which consists of verifying extracted term pairs with access to sample contexts as well as statistical data. In IView all token alignments made by ITrix are compiled into a table of translation pair types in a graphical environment where the annotator can confirm translation pairs as domain specific terms or as belonging to “general language”, i.e., they are correct alignments but cannot be considered as terms in a specific domain.

3.2 Metrics

As hinted in the introduction, it is desirable to optimize the quality of the aligned data by stripping away poor quality alignments and keeping the high quality ones as this will leave less manual work in the actual standardization process. To achieve this, one needs to order the proposed term pairs in, for example, descending quality order. Ordering term candidates can be done using different metrics.

One such metric that has been used in term extraction research is the Dice’s coefficient of association (Dice 1945). A common approach in applying Dice’s coefficient as a ranking metric is to collect corpus statistics (Pazienza et al. 2005). The second metric used in this study is the Q-value, a metric specifically design to operate on aligned data (Deléger et al. 2006). These two metrics are compared to a third baseline, which is a straightforward pair frequency.

The input data used for these metrics are all available in the SQL database, which contains information such as

- **T**ype **P**air **F**requencies (TPF), i.e. the number of times where the source and target types are aligned
- **T**arget types **p**er **S**ource type (TpS), i.e. the number of target types a specific source type has been aligned to. E.g. if the source type A is aligned to the target types B and C, two type pairs exist – A-B and A-C. For both these type pairs, the TpS value is 2.
- **S**ource types **p**er **T**arget type (SpT), i.e. the number of source types a specific target type has been aligned to. Given the example provided to explain the TpS, the SpT values for the two type pairs would be 1 for A-B, and 1 for A-C. This means that low SpT and TpS values correspond to consistent usage of target and source types if the aligned data is fairly correct.
- **S**ource **T**ype **F**requency (STF), i.e. the accumulated frequency of a source type in the set of aligned type pairs.
- **T**arget **T**ype **F**requency (STF), i.e. the accumulated frequency of a target type in the set of aligned type pairs.

Using this information, we can calculate the following metrics:

$$Q-value = \frac{TPF}{TpS + SpT}$$

$$Dice = \frac{2 \times TPF}{STF + TTF}$$

#	Src	Trg	TPF	TpS	SpT	STF	TTF
1	fatty acid	Fett syra	2	2	1	7	2
2	fatty acid	fettsyra	5	2	1	7	5

Table 1. Two type pairs and their frequencies.

Given the complete set of type pairs in Table 1, the Q-value of pair 1 is 0.67 and the Dice coefficient is 0.45. The Q-value of pair 2 is 0.71 and the Dice coefficient is 0.83.

The main conceptual difference between the Dice coefficient and the Q-value is that the Dice coefficient focuses on positive association between source and target type, whereas the Q-value focuses on the association between the current source and target type, but also between the current source and target types with other source and target types.

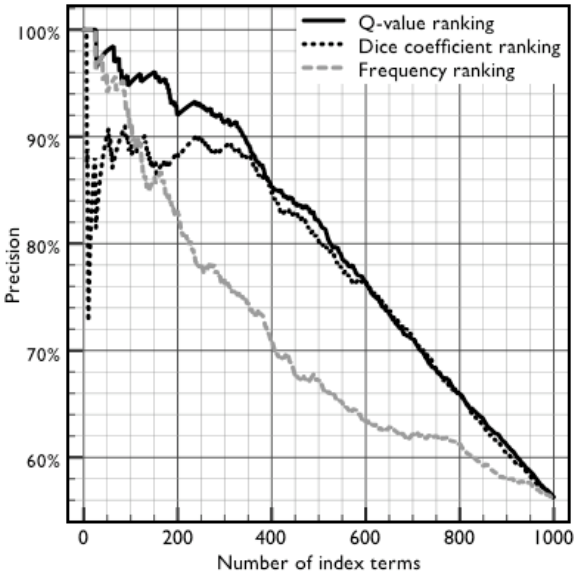

Figure 2. Precision fall-off using different metrics to rank 1000 randomly sampled term candidates

In other words, a high Q-value indicates a term candidate pair with few similar candidates whereas a high Dice coefficient indicates a common term.

4 Evaluation and results

Processing the patent texts in the ITools suite resulted in over 60,000 term candidate pairs in the term bank. One thousand entries were sampled randomly from these term pairs and evaluated manually for correctness. The manually corrected test set of term candidates was then ordered by using the three different metrics: term pair frequency, Dice coefficient, and Q-value. The results are presented in Figure 2.

As can be seen in Figure 2, ranking the term candidates using the Q-value results in the best accumulated precision curve. Both Q-value and Dice coefficient metrics rank the term candidates in a fairly linear correlation with term candidate precision, whereas the term pair frequency curve has a bad precision fit. The Dice coefficient does not perform well at its highest scores, which could be explained by the fact that the term pairs contain a considerable amount of term pairs where frequency is equal to one (1) (over 50 per cent of the term pairs).

5 Discussion

As stated earlier, the motivation of finding a good term ranking metric is to increase the efficiency of the validation process – the process of going from term candidates to standardized terms. If we assume that the random sample is representative of the total 60,000 term candidates generated by ITools, we can choose a combination of precision and coverage by setting a threshold at the appropriate Q value.

Precision	Q-value	Number of term candidates to process	Estimated number of approved terms
95.8%	0.53	10,400	9963
91.0%	0.50	20,040	18236
~80.0%	0.20	~30,000	~24,000

Table 2. Estimated term volumes and precision for different Q-values.

In table 2, three different Q-values have been chosen resulting in three sets of term candidates to process. Set 2 is double the size of set 1, and set 3 is three times the size of set 1. The increases in size can roughly be translated into the same increase in time needed to process the term candidates. The precision of the sets gives us an estimate of how many approved terms we can expect from processing a given number of term candidates. Furthermore, given a scenario where there are no resource restrictions enforced on the validation process, processing the full set of term candidates will of course result in the highest number of approved terms.

However, resources available for the validation process are often limited. In this case the precision of the set of term candidates becomes interesting as this can roughly be translated into processing efficiency. If we assume that all term candidates require the same amount of processing time, we can use the data in Table 1 to derive the earnings and costs connected to the different sizes of term candidate sets. An example of such calculations is presented in Table 3.

Increase in term candidate volume	Additional effort required	Additional approved terms gained	Difference between effort and gain
10400-20040	92.7%	83%	-9.7
20040-30000	49,7%	31.6%	-18.1

Table 3. Earnings and costs when increasing the volume of term candidates to process.

Using the three sets of term candidates presented in Table 2, the relative increases in effort (spent time) and number of approved terms, as well as the difference between these gains and efforts have been calculated in Table 3. As we can see, an additional increase in candidate terms from 20 000 to 30 000 results in half the effective gain of the resources spent, compared to an increase in term candidates to process from 10 000 to 20 000. In effect, having a good term ranking metric with a predictable precision fall rate can provide the information necessary to come to the decision on how term processing resources can be spent in the most effective way, depending on the present requirements on the size of the final set of approved terms.

6 Conclusion and future work

In this paper we have shown that time and effort could be saved by a relatively simple evaluation metric based on frequency data from term pairs, source and target distributions inside the alignment results. The proposed metric Q-value is shown to outperform other tested metrics such as the Dice coefficient, and simple pair frequency. The Q-value is better at handling low frequency data. The results point to that one could realistically decide on what goals a term standardization endeavor should aim for in terms of volume, and time spent.

The next step to develop the methodology for revising term candidates further is to test it on several alignment projects. We are currently investigating techniques to cluster term pairs conceptually by using semantic mirroring using a Q-value filter. The initial results look promising in that they make it possible to group synonym variants within a conceptual cluster and thereby making it possible to automatically filter out undesired term synonyms.

References

Ahrenberg L, Merkel M, Petterstedt M. 2003. Interactive Word Alignment for Language Engineering. In *Proceedings of the 10th Conference of the EACL*, Budapest.

Deléger L, Merkel M, Zweigenbaum P. 2006. Enriching Medical Terminologies: an Approach Based on Aligned Corpora. In the *Proceedings 20th International Congress of the European Federation for Medical Informatics (MIE 2006)*.

Dice, L R. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, Vol 26, pp 297-302.

Foo J, Merkel M. 2007. Building standardized term banks through automated term extraction and advanced editing tools. To be published in the *Proceedings from the International Conference on Terminology* 2006, November 16-17 Antwerp.

Och F, Ney H. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, vol 29, number 1, pp. 19-51.

Lombard R. 2006. Managing source-language terminology. In K. J. Dunne (ed.), *Perspectives on Localization.* John Benjamins Publishing Company. Amsterdam.

Nyström M, Merkel M, Ahrenberg L, Zweigenbaum P, Petersson H, and Åhlfeldt H. 2006. Creating a medical English-Swedish dictionary using interactive word alignment. In *BMC Medical Informatics and Decision Making 2006*, 6:35.

Pazienza M. T, Pennacchiotti M, Zanzotto F M. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. In: S. Sirmakessis (ed.) *Knowledge Mining*. Series: Studies in Fuzziness and Soft Computing, Vol.185, Springer Verlag.

Tapanainen P and Järvinen T. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing: 31 March-3 April 1997; Washington D.C* , pp. 64-71.

Tiedemann J. 2003. Combining clues for word alignment. In *Proceedings of the 10th Conference of the EACL*, Budapest

Wright, S. E. 2006. *The role of terminology management in Localization*. Teminology seminar given on Internet (webinar), SDL, May 20, 2006.