Using Weighted Abduction to Align Term Variant Translations in Bilingual Texts

Michael Carl, Ecaterina Rascu and Johann Haller

Institut für Angewandte Informationsforschung
Martin-Luther-Str. 14, Saarbrücken, Germany
{carl, kati, hans}@iai.uni-sb.de

Abstract

In this paper we describe a method for detecting terminological variants and their translations in bilingual texts. Our approach is based on abductive reasoning and combines various monolingual and bilingual resources. A small scale experiment shows that precision and recall increase when using more resources and when the resources interfere in a less restricted way. In order to tune our system, we develop a weighing strategy based on the precision of term translation alignments in a reference text. We feed these weights back into the linguistic resources and repeat the experiment. The results show that precision values are considerably higher when weighing term alignments.

1. Introduction

The consistent use of terms in technical domains increases the comprehensibility and translatability of texts (Mitamura and Nyberg, 1995). However, terminological variation is a frequent phenomenon even in established domains (Daille et al., 1996; Macklovitch, 1995; Royauté, 1999).

Enguehardt distinguishes between term recognition systems (TRS) and term extraction systems (Enguehard, 2003). While the latter identify new terms in texts, the former ones detect variants of already known terms.

In this paper we investigate an abductive method to detect terms and their translations in bilingual texts. The proposed architecture combines two monolingual term recognition systems capable of identifying terms and their variants. We infer term variation templates from language specific general variation patterns by means of abduction and use them to identify term variant translations in aligned texts. Abduction as "inference to the best explanation" also requires a ranking of the hypotheses by evaluating their explanatory power (Magnani, 2001). We achieve this by weighing term variation templates according to the co-occurrence precision of variation patterns.

We first present the approach adopted for term recognition. In section 3., we evaluate the system in a number of different settings.

2. Abductive Approach to Term Recognition

To detect translations of terms and their variants in an aligned English–French text, the system requires two types of resources. The first resource is a bilingual terminology containing base terms and their authorized translations. The second resource consists of language specific general variation patterns and synonymy relations. Based on these variation patterns and the terminology, a number of term specific variation templates are generated for every term in the bilingual terminology. The variation templates are stored in a database — the so-called Abductive Terminology Database (ATDB) — together with the original terms so that each variant is linked to its authorized form. The architecture is plotted in figure 1. The actual ATDB is in the center of figure 1 and will be discussed in section 2.2..

An ATDB consists of two symmetrical language sides, a left-hand English side and a right-hand French side[1]. A bilingual sentence aligned text is fed into the system which detects term translations and marks them accordingly. The automatically annotated text is then compared with the manually annotated version of the same text. Values for precision and recall are computed for every term and template. We accumulate weights for general variation patterns based on precision values of term templates and feed these weights back into the resources. We refer to this mechanism as weighted abduction. In section 3. we outline this approach in more depth and show how it can be used to grade ambiguities and reduce noise.

In the following subsections we present the different resources of the ATDB in more detail.

2.1. General Term Variation Patterns

We distinguish three types of variations: typographical variations, morpho-syntactic variations and lexical variations. In this section we give examples of these.

2.1.1. Typographical Variation

Typographical variants differ in the way hyphenation, blanks or punctuation marks are used around a term constituent. Examples are given in (1) and (2). We write the authorized term on the left-hand side of the arrow and the variant on the right-hand side.

(1) *hand stop* → *handstop*
(2) *re-insert* → *reinsert*

2.1.2. Morpho-syntactic Variation

Morpho-syntactic variants are derived from a base term by morphological derivation and/or by transformation of its syntactic structure. The basic mechanisms of structural transformation are omission, insertion, permutation, and coordination (Jacquemin, 1996; Daille et al., 1996). Omission implies the deletion of one or more components from

[1] see also (Carl et al., 2004) for a more detailed discussion.

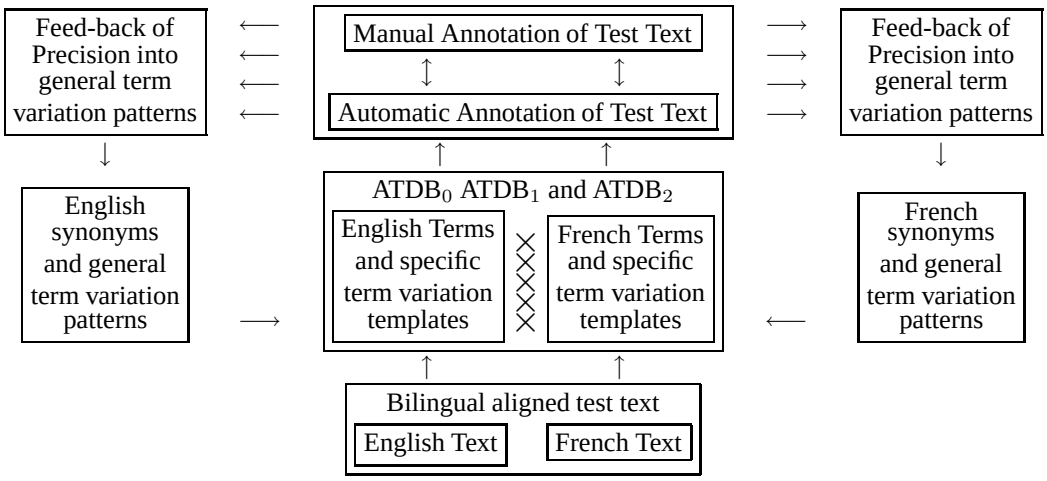

Figure 1: Architecture of the ATDB

a multi-word unit producing variants that are more generic than the original base term (Jacquemin, 1996). By inserting one or more elements into a base term, more specific variants are produced. Variation by permutation changes the linear order of the constituents in a base-term (Daille et al., 1996). Examples (3) and (4) show English and French permutation variants respectively. The lower line in these variants show a generalized variation pattern[2].

(3) *rifle butt* → *butt of a rifle*
N$_1$ N$_2$ → N$_2$ p d N$_1$

(4) fusil de tireur d'élite c3a1 →
fusil c3a1 de le tireur d'élite
N$_1$ p$_2$ N$_3$ p$_4$^ d$_5$^ N$_6$ N$_7$ →
N$_1$ N$_7$ p$_2$ d N$_3$ p$_4$^ d$_5$^ N$_6$

2.1.3. Lexical Variation

Terms also vary in the choice of their lexemes. In order to detect such variants we consider synonymy relations. A problem related to synonyms is to determine a sufficient context, so that unnecessary noise is avoided and a maximum number of relevant terms are found. Synonyms are highly domain specific (Carl et al., 2002). This implies that for restricted domains we can shorten the context in which a synonym occurs without risking to produce too much noise. The synonyms in (5) and (6) are obtained by substituting the French modifier visée → tir and the English head word *telescope* → *scope*.

(5) *spotting telescope* → *spotting scope*
(6) lunette de visée → lunette de tir

[2]The general patterns describe the variation mechanism using part-of-speech tags, N, A, d, and p for noun, adjective, determiner, and preposition respectively. The sign "^" expresses optionality of the preceeding tag. Indexed tags map the word from left to right, non-indexed tags are inserted in or deleted from the variant.

2.2. Specific Term Variation Templates

Starting from an initial database of unambiguous term translations, English and French variant templates are abduced using synonym lists and general variation patterns. We describe this process by means of an example. Assume, for instance, the term translations (7) *spotting telescope* ↔ lunette d' observation and (8) *telescopic sight* ↔ lunette de visée in table 1 are contained in the base terminology. A number of variants and variant templates can be abduced from these terms using general variation patterns. The pattern names are given in the first column for English and in the last column for French in table 1. By inspecting English and French texts, we have induced 13 English and 16 French general variation patterns for omission, insertion, permutation, coordination and synonymy. Some of these patterns are shown in table 1; for a more detailed description see (Carl et al., 2004). From term (7) the variant *telescope* is obtained by applying omission pattern EO_1 while the variant *scope* is the result of successively applying the synonymy pattern ES_1 and the variation pattern EO_1.

By taking into account various combinations of resources, we generate three different databases. The database ATDB$_0$ is identical to the original term database, with no additional variants. ATDB$_1$ includes ATDB$_0$ as well as all variants that can be generated through a single application of one variation or synonymy pattern. Five variants are produced for *spotting telescope* and three variants are produced for *telescopic sight* in ATDB$_1$. In addition to the entries in ATDB$_1$, ATDB$_2$ contains variants in which synonyms co-occur with variation patterns. This produces another four variants. Table 1 contains three sections separated through horizontal lines. These sections represent the abduced variation templates for the databases ATDB$_0$, ATDB$_1$, and ATDB$_2$ respectively. Note that ATDB$_2$ does not contain additional entries for lunette d'observation and for *telescopic sight*, due to lack of appropriate synonyms.

While the original terminology (and correspondingly ATDB$_0$) contains only $1 - to - 1$ term translation correspondences, ATDB$_1$ and ATDB$_2$ contain $m - to - n$ translation relations. This generates term translation ambigu-

(7) *spotting telescope* ↔ lunette d' observation

Pattern	English ATDB		French ATDB	Pattern
$base$	*spotting telescope*			
EO_1	*telescope*			
EI_1	*spotting* A *telescope*			
EP_3	*telescope* p d *spotting*			
ECO_1	*spotting* Conj (N;A) *telescope*	↔	lunette d' observation	$base$
ES_1	*spotting scope*		lunette d' d observation A	FI_2
ES_1, EO_1	*scope*		lunette d N Conj d'observation	FCO_1
ES_1, EI_1	*spotting* A *scope*		lunette	FO_1
ES_1, EP_3	*scope* p d *spotting*			
ES_1, ECO_1	*spotting* Conj (N;A) *scope*			

(8) *telescopic sight* ↔ lunette de visée

Pattern	English ATDB		French ATDB	Pattern
			lunette de visée	$base$
$base$	*telescopic sight*		lunette de d visée A	FI_2
EI_1	*telescopic* A *sight*		lunette d N Conj de visée	FCO_1
ECO_1	*telescopic* Conj (N;A) *sight*	↔	lunette	FO_1
EO_1	*sight*		lunette de tir	FS_1
			lunette d N Conj de tir	FS_1, FCO_1
			lunette	FS_1, FO_1

Table 1: Abduction of Term Variants

ities. Due to variation pattern FO_1, French lunette, for instance, is recognized as an omission variant of the base terms lunette de visée and lunette d'observation. We expect that adding further terms and variation patterns to the resources increases the ambiguity of the terminology. Ambiguity is also likely to increase for higher level ATDB$_i$, $i > 2$. As we show in the next section, coverage and precision also increase with higher level ATDBs.

3. Experiments and Evaluation

In this section we evaluate the performance of the ATDBs. An overall picture of the evaluation architecture is shown in figure 1. We use two bilingual aligned test texts, SNIPER2 and SNIPER3, two excerpts from an army manual on sniper training and deployment (Macklovitch, 1995). SNIPER2 and SNIPER3 have 391 and 400 English–French aligned segments, respectively, with an average length of 19 and 22 words in the English and the French segment. For the evaluation of the ATDBs we established "gold standards" by manually annotating term translations in SNIPER2 and SNIPER3.

As outlined in section 2., the abduction of term variants in the ATDB requires a bilingual terminology. The bilingual terminology used for the evaluation of the ATDB was manually extracted from the test texts. It contains 154 non-ambiguous term translations where each English and French term occurs exactly once. Two small sets of 131 synonyms for 54 English content words and 92 synonyms for 50 French content words were also collected manually. We induced 13 general variation patterns for English and 16 for French terms from the same texts. These are presented in more detail in (Carl et al., 2004).

We generate three databases ATDB$_0$, ATDB$_1$ and ATDB$_2$ from these resources and the general variation patterns as described in section 2.2.. While ATDB$_0$ contains only 154 base terms, ATDB$_1$ contains in addition 318 term variation templates for English and 508 variation templates for French. ATDB$_2$ has 461 English and 699 French variation templates.

The test texts are passed through ATDB$_0$, ATDB$_1$, and ATDB$_2$ where terms and their variants are marked automatically. These results are compared with the manual annotation of the texts (see figure 1) and values of precision and recall are computed for the three databases. Table 2 summarizes the results.

	SNIPER2		
	ATDB$_0$	ATDB$_1$	ATDB$_2$
precision	0.53	0.61	0.62
recall	0.45	0.78	0.89
correct	467	802	916
noise	407	510	562
misses	566	231	117
precision$_w$		0.79	0.77
	SNIPER3		
precision	0.59	0.64	0.66
recall	0.40	0.79	0.86
correct	373	732	804
noise	259	410	422
misses	557	198	126
precision$_w$		0.81	0.79

Table 2: Coverage and Precision of the ATDB

Table 2 shows that when using more resources (ATDB$_1$ and ATDB$_2$), precision *and* recall increase. Increase in recall is, however, much more significant than the increase in precision. The high amount of noise produced is due to the

following reasons:

- For $ATDB_0$, noise is mainly due to the fact that terms were detected in one side of the alignment but no corresponding translation could be found in the other, i.e. these terms are translated through a variant in the text.
- Ambiguities also occur when aligning under-specified variants. For instance, in case the English segment contains the word *position* and $ATDB_1$ and $ATDB_2$ do not know to which of the three base terms *firing position*, *hawking position*, or *prone position* it refers. A word can thus be detected as a variant of several base terms.
- Noise also occurs when establishing all possible connections between pairs of terms (and their variants) in both language sides. For instance, in cases where the word lunette occurs twice or more often in a French sentence and the word *scope*, *telescope* or *sight* etc. occurs in the English sentence, each occurrence of lunette is aligned with every occurrence of the English variant.

As the number of resources increases and the way they interact multiplies, alignment prediction becomes more ambiguous. This can be seen in the noise produced for higher order ATDBs. Performing a syntactic analysis would allow to avoid some of these multiple connections between terms and reduce noise. Our estimations show that precision could increase by ca. 20% with an appropriate syntactic analysis. Another method to reduce ambiguous connections — which we will follow in the rest of this section — is to rank term alignments and predict which of the possible links is most reliable.

Based on the precision of term alignment for SNIPER2 as shown in table 2 we compute weights for base terms, general variation patterns and synonyms. The weights are calculated as the co-occurrence precision for every three-tuple which consists of a base term translation, an English variation pattern and a French variation pattern. In the second run, these weights are associated with the variation templates in order to rank the strength of two or more ambiguous term alignments. Computing precision of this second run generates the figures for precision as shown in the last line in table 2. The gain in weighted precision $precision_w$ compared to the non-weighted *precision* is higher in $ATDB_1$ than in $ATDB_2$ for the two texts. However, it is higher than 12% in both ATDBs. Thus, by feeding evaluation values back into the abduction process, we enhance the precision of the tool considerably.

From the higher precision obtained through weighted term alignments we conclude that certain variation types tend to co-occur more frequently than others. We believe that this insight can be of great value and should be exploited more thoroughly in the further development of the tool.

4. Conclusions

This paper presents an Abductive Terminology Database (ATDB), a tool designed to detect term translations and their variants in aligned bilingual texts. The tool integrates different resources, a terminology database, lists of synonyms and sets of general variation patterns, which can be combined in various ways. As the number of resources increases and the way they interact multiplies, alignment of term translations becomes more ambiguous. However, we find at the same time that recall and precision also increase in most cases. To further enhance precision we have implemented and discussed a method to weigh and rank term alignments. We tune the ATDB by feeding these values back into the general variation patterns and synonym lists. In our opinion this provides a powerful means not only to enhance alignment results but also to investigate on an empirical basis the success and applicability of variation patterns in different contexts.

5. References

Carl, Michael, Johann Haller, Christoph Horschmann, Dieter Maas, and Jörg Schütz, 2002. The TETRIS Terminology Tool. *TAL, Structuration de terminologie*, 43(1).

Carl, Michael, Ecaterina Rascu, Johann Haller, and Philippe Langlais, 2004. Abducing Term Variant Translations in Aligned Texts. *Terminology*, to appear.

Daille, Béatrice, Benoît Habert, Christian Jacquemin, and Jean Royauté, 1996. Empirical Observation of Term Variations and Principles for their Description. *Terminology*, 3(2).

Enguehard, Chantal, 2003. CoRRecT : Démarche coopérative pour l'évaluation de systèmes de reconnaissance de termes. In *in Proceedings of TALN*.

Jacquemin, Christian, 1996. A symbolic and surgical acquisition of terms through variation. In *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*.

Macklovitch, Elliott, 1995. Can terminological consistency be validated automatically? Technical report, CITI/RALI, Montréal, Canada.

Magnani, Lorenzo, 2001. *Abduction, Reason and Science. Processes of Discovery and Explanation*. Doordrecht: Kluwer Academic.

Mitamura, Teruko and Eric Nyberg, 1995. Controlled English for Knowledge-Based MT: Experience with the KANT System. In *Proceedings of TMI-95*.

Royauté, Jean, 1999. *Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information*. Ph.D. thesis, Université Henri Poincaré-Nancy 1 College, Nancy.