Language-independent bilingual terminology extraction from a multilingual parallel corpus

Els Lefever[1,2], **Lieve Macken**[1,2] **and Veronique Hoste**[1,2]

[1]LT3
School of Translation Studies
University College Ghent
Groot-Brittanniëlaan 45
9000 Gent, Belgium

[2]Department of Applied Mathematics
and Computer Science
Ghent University
Krijgslaan281-S9
9000 Gent, Belgium

{Els.Lefever, Lieve.Macken, Veronique.Hoste}@hogent.be

Abstract

We present a language-pair independent terminology extraction module that is based on a sub-sentential alignment system that links linguistically motivated phrases in parallel texts. Statistical filters are applied on the bilingual list of candidate terms that is extracted from the alignment output.

We compare the performance of both the alignment and terminology extraction module for three different language pairs (French-English, French-Italian and French-Dutch) and highlight language-pair specific problems (e.g. different compounding strategy in French and Dutch).

Comparisons with standard terminology extraction programs show an improvement of up to 20% for bilingual terminology extraction and competitive results (85% to 90% accuracy) for monolingual terminology extraction, and reveal that the linguistically based alignment module is particularly well suited for the extraction of complex multiword terms.

1 Introduction

Automatic Term Recognition (ATR) systems are usually categorized into two main families. On the one hand, the linguistically-based or rule-based approaches use linguistic information such as PoS tags, chunk information, etc. to filter out stop words and restrict candidate terms to predefined syntactic patterns (Ananiadou, 1994), (Dagan and Church, 1994). On the other hand, the statistical corpus-based approaches select n-gram sequences as candidate terms that are filtered by means of statistical measures. More recent ATR systems use hybrid approaches that combine both linguistic and statistical information (Frantzi and Ananiadou, 1999).

Most bilingual terminology extraction systems first identify candidate terms in the source language based on predefined source patterns, and then select translation candidates for these terms in the target language (Kupiec, 1993).

We present an alternative approach that generates candidate terms directly from the aligned words and phrases in our parallel corpus. In a second step, we use frequency information of a general purpose corpus and the n-gram frequencies of the automotive corpus to determine the term specificity. Our approach is more flexible in the sense that we do not first generate candidate terms based on language-dependent predefined PoS patterns (e.g. for French, N N, N Prep N, and N Adj are typical patterns), but immediately link linguistically motivated phrases in our parallel corpus based on lexical correspondences and syntactic similarity.

This article reports on the term extraction experiments for 3 language pairs, i.e. French-Dutch, French-English and French-Italian. The focus was on the extraction of automative lexicons.

The remainder of this paper is organized as follows: Section 2 describes the corpus. In Section 3 we present our linguistically-based sub-sentential alignment system and in Section 4 we describe how we generate and filter our list of candidate terms. We compare the performance of our system with both bilingual and monolingual state-of-the-art terminology extraction systems. Section 5 concludes this paper.

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 496–504,
Athens, Greece, 30 March – 3 April 2009.

2 Corpus

The focus of this research project was on the automatic extraction of 20 bilingual automative lexicons. All work was carried out in the framework of a customer project for a major French automotive company. The final goal of the project is to improve vocabulary consistency in technical texts across the 20 languages in the customer's portfolio. The French database contains about 400,000 entries (i.e. sentences and parts of sentences with an average length of 9 words) and the translation percentage of the database into 19 languages depends on the target market.

For the development of the alignment and terminology extraction module, we created three parallel corpora (Italian, English, Dutch) with French as a central language. Figures about the size of each parallel corpus can be found in table 1.

	Target Lang.	# Sentence pairs	# words
French	Italian	364,221	6,408,693
French	English	363,651	7,305,151
French	Dutch	364,311	7,100,585

Table 1: Number of sentence pairs and total number of words in the three parallel corpora

2.1 Preprocessing

We PoS-tagged and lemmatized the French, English and Italian corpora with the freely available TreeTagger tool (Schmid, 1994) and we used TadPole (Van den Bosch et al., 2007) to annotate the Dutch corpus.

In a next step, chunk information was added by a rule-based language-independent chunker (Macken et al., 2008) that contains distituency rules, which implies that chunk boundaries are added between two PoS codes that cannot occur in the same constituent.

2.2 Test and development corpus

As we presume that sentence length has an impact on the alignment performance, and thus on term extraction, we created three test sets with varying sentence lengths. We distinguished short sentences (2-7 words), medium-length sentences (8-19 words) and long sentences ($>$ 19 words). Each test corpus contains approximately 9,000 words; the number of sentence pairs per test set can be found in table 2. We also created a development corpus with sentences of varying length to debug the linguistic processing and the alignment module as well as to define the thresholds for the statistical filtering of the candidate terms (see 4.1).

	# Words	# Sentence pairs
Short ($<$ 8 words)	+- 9,000	823
Medium (8-19 words)	+- 9,000	386
Long ($>$ 19 words)	+- 9,000	180
Development corpus	+-5,000	393

Table 2: Number of words and sentence pairs in the test and development corpora

3 Sub-sentential alignment module

As the basis for our terminology extraction system, we used the sub-sentential alignment system of (Macken and Daelemans, 2009) that links linguistically motivated phrases in parallel texts based on lexical correspondences and syntactic similarity. In the first phase of this system, *anchor chunks* are linked, i.e. chunks that can be linked with a very high precision. We think these anchor chunks offer a valid and language-independent alternative to identify candidate terms based on predefined PoS patterns. As the automotive corpus contains rather literal translations, we expect that a high percentage of anchor chunks can be retrieved.

Although the architecture of the sub-sentential alignment system is language-independent, some language-specific resources are used. First, a bilingual lexicon to generate the lexical correspondences and second, tools to generate additional linguistic information (PoS tagger, lemmatizer and a chunker). The sub-sentential alignment system takes as input sentence-aligned texts, together with the additional linguistic annotations for the source and the target texts.

The source and target sentences are divided into chunks based on PoS information, and lexical correspondences are retrieved from a bilingual dictionary. In order to extract bilingual dictionaries from the three parallel corpora, we used the Perl implementation of IBM Model One that is part of the Microsoft Bilingual Sentence Aligner (Moore, 2002).

In order to link chunks based on lexical clues and chunk similarity, the following steps are taken for each sentence pair:

1. Creation of the lexical link matrix
2. Linking chunks based on lexical correspondences and chunk similarity

3. Linking remaining chunks

3.1 Lexical Link Matrix

For each source and target word, all translations for the word form and the lemma are retrieved from the bilingual dictionary. In the process of building the lexical link matrix, function words are neglected. For all content words, a lexical link is created if a source word occurs in the set of possible translations of a target word, or if a target word occurs in the set of possible translations of the source words. Identical strings in source and target language are also linked.

3.2 Linking Anchor chunks

Candidate anchor chunks are selected based on the information available in the lexical link matrix. The candidate target chunk is built by concatenating all target chunks from a *begin index* until an *end index*. The begin index points to the first target chunk with a lexical link to the source chunk under consideration. The end index points to the last target chunk with a lexical link to the source chunk under consideration. This way, 1:1 and 1:n candidate target chunks are built. The process of selecting candidate chunks as described above, is performed a second time starting from the target sentence. This way, additional n:1 candidates are constructed. For each selected candidate pair, a *similarity test* is performed. Chunks are considered to be similar if at least a certain percentage of words of source and target chunk(s) are either linked by means of a lexical link or can be linked on the basis of corresponding part-of-speech codes. The percentage of words that have to be linked was empirically set at 85%.

3.3 Linking Remaining Chunks

In a second step, chunks consisting of one function word – mostly punctuation marks and conjunctions – are linked based on corresponding part-of-speech codes if their left or right neighbour on the diagonal is an anchor chunk. Corresponding final punctuation marks are also linked.

In a final step, additional candidates are constructed by selecting non-anchor chunks in the source and target sentence that have corresponding left and right anchor chunks as neigbours. The anchor chunks of the first step are used as contextual information to link n:m chunks or chunks for which no lexical link was found in the lexical link matrix.

In Figure 1, the chunks [Fr: gradient] – [En: gradient] and the final punctuation mark have been retrieved in the first step as anchor chunk. In the last step, the n:m chunk [Fr: de remontée pédale d' embrayage] – [En: of rising of the clutch pedal] is selected as candidate anchor chunk because it is enclosed within anchor chunks.

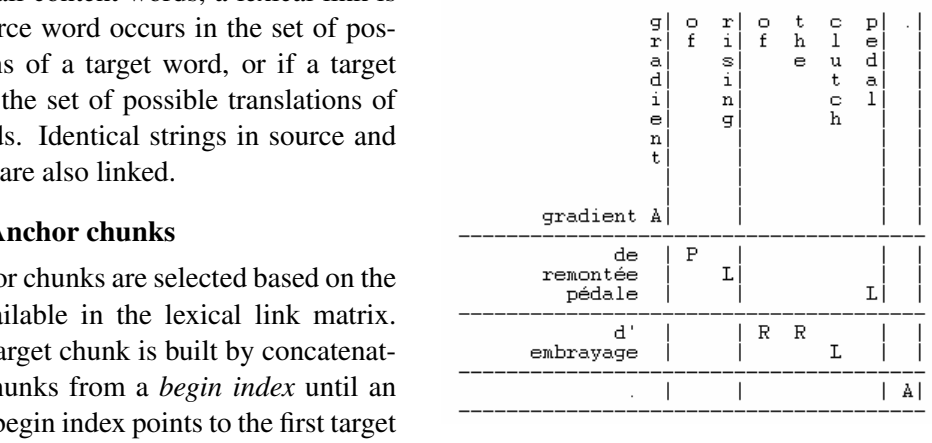

Figure 1: n:m candidate chunk: 'A' stands for anchor chunks, 'L' for lexical links, 'P' for words linked on the basis of corresponding PoS codes and 'R' for words linked by language-dependent rules.

As the contextual clues (the left and right neigbours of the additional candidate chunks are anchor chunks) provide some extra indication that the chunks can be linked, the similarity test for the final candidates was somewhat relaxed: the percentage of words that have to be linked was lowered to 0.80 and a more relaxed PoS matching function was used.

3.4 Evaluation

To test our alignment module, we manually indicated all translational correspondences in the three test corpora. We used the evaluation methodology of Och and Ney (2003) to evaluate the system's performance. They distinguished *sure* alignments (S) and *possible* alignments (P) and introduced the following redefined precision and recall measures (where A refers to the set of alignments):

$$precision = \frac{|A \cap P|}{|A|}, recall = \frac{|A \cap S|}{|S|} \quad (1)$$

and the alignment error rate (AER):

$$AER(S, P; A) = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|} \quad (2)$$

Table 3 shows the alignment results for the three language pairs. (Macken et al., 2008) showed that the results for French-English were competitive to state-of-the-art alignment systems.

	SHORT			MEDIUM			LONG		
	p	r	e	p	r	e	p	r	e
Italian	.99	.93	.04	.95	.89	.08	.95	.89	.07
English	.97	.91	.06	.95	.85	.10	.92	.85	.12
Dutch	.96	.83	.11	.87	.73	.20	.87	.67	.24

Table 3: Precision (p), recall (r) and alignment error rate (e) for our sub-sentential alignment system evaluated on French-Italian, French-English and French-Dutch

As expected, the results show that the alignment quality is closely related to the similarity between languages. As shown in example (1), Italian and French are syntactically almost identical – and hence easier to align, English and French are still close but show some differences (e.g different compounding strategy and word order) and French and Dutch present a very different language structure (e.g. in Dutch the different compound parts are not separated by spaces, *separable verbs*, i.e. verbs with prefixes that are stripped off, occur frequently (*losmaken* as an infinitive versus *maak los* in the conjugated forms) and a different word order is adopted).

(1) Fr: déclipper le renvoi de ceinture de sécurité.
(En: unclip the mounting of the belt of safety)
It: sganciare il dispositivo di riavvolgimento della cintura di sicurezza.
(En: unclip the mounting of the belt of satefy)
En: unclip the seat belt mounting.
Du: maak de oprolautomaat van de autogordel los.
(En: clip the mounting of the seat-belt un)

We tried to improve the low recall for French-Dutch by adding a decompounding module to our alignment system. In case the target word does not have a lexical correspondence in the source sentence, we decompose the Dutch word into its meaningful parts and look for translations of the compound parts. This implies that, without decompounding, in example 2 only the correspondences *doublure – binnenpaneel*, *arc – dakversteviging* and *arrière – achter* will be found. By decomposing the compound into its meaningful parts (binnenpaneel = binnen + paneel, dakversteviging = dak + versteviging) and retrieving the lexical links for the compound parts, we were able to link the missing correspondence: *pavillon – dakversteviging*.

(2) Fr: doublure arc pavillon arrière.
(En: rear roof arch lining)
Du: binnenpaneel dakversteviging achter.

We experimented with the decompounding module of (Vandeghinste, 2008), which is based on the Celex lexical database (Baayen et al., 1993). The module, however, did not adapt well to the highly technical automotive domain, which is reflected by its low recall and the low confidence values for many technical terms. In order to adapt the module to the automotive domain, we implemented a domain-dependent extension to the decompounding module on the basis of the development corpus. This was done by first running the decompounding module on the Dutch sentences to construct a list with possible *compound heads*, being valid compound parts in Dutch. This list was updated by inspecting the decompounding results on the development corpus. While decomposing, we go from right to left and strip off the longest valid part that occurs in our preconstructed list with compound parts and we repeat this process on the remaining part of the word until we reach the beginning of the word.

Table 4 shows the impact of the decompounding module, which is more prominent for short and medium sentences than for long sentences. A superficial error analysis revealed that long sentences combine a lot of other French – Dutch alignment difficulties next to the decompounding problem (e.g. different word order and separable verbs).

	SHORT			MEDIUM			LONG		
	p	r	e	p	r	e	p	r	e
Dutch									
no_dec	.95	.76	.16	.88	.67	.24	.88	.64	.26
dec	.96	.83	.11	.87	.73	.20	.87	.67	.24

Table 4: Precision (p), recall (r) and alignment error rate (e) for French-Dutch without and with decompounding information

4 Term extraction module

As described in Section 1, we generate candidate terms from the aligned phrases. We believe these anchor chunks offer a more flexible approach

because the method is language-pair independent and is not restricted to a predefined set of PoS patterns to identify valid candidate terms. In a second step, we use a general-purpose corpus and the n-gram frequency of the automotive corpus to determine the specificity of the candidate terms.

The candidate terms are generated in several steps, as illustrated below for example (3).

(3) Fr: Tableau de commande de climatisation automatique
En: Automatic air conditioning control panel

1. Selection of all anchor chunks (minimal chunks that could be linked together) and lexical links within the anchor chunks:

tableau de commande	control panel
climatisation	air conditioning
commande	control
tableau	panel

2. combine each NP + PP chunk:

commande de climatisation automatique	automatic air conditioning control
tableau de commande de climatisation automatique	automatic air conditioning control panel

3. strip off the adjectives from the anchor chunks:

commande de climatisation	air conditioning control
tableau de commande de climatisation	air conditioning control panel

4.1 Filtering candidate terms

To filter our candidate terms, we keep following criteria in mind:

- each entry in the extracted lexicon should refer to an object or action that is relevant for the domain (notion of *termhood* that is used to express "the degree to which a linguistic unit is related to domain-specific context" (Kageura and Umino, 1996))
- multiword terms should present a high degree of cohesiveness (notion of *unithood* that expresses the "degree of strength or stability of syntagmatic combinations or collocations" (Kageura and Umino, 1996))
- all term pairs should contain valid translation pairs (translation quality is also taken into consideration)

To measure the termhood criterion and to filter out general vocabulary words, we applied Log-Likelihood filters on the French single-word terms. In order to filter on low unithood values, we calculated the Mutual Expectation Measure for the multiword terms in both source and target language.

4.1.1 Log-Likelihood Measure

The Log-Likehood measure (LL) should allow us to detect single word terms that are *distinctive* enough to be kept in our bilingual lexicon (Daille, 1995). This metric considers word frequencies weighted over two different corpora (in our case a technical automotive corpus and the more general purpose corpus "Le Monde"[1]), in order to assign high LL-values to words having much higher or lower frequencies than expected. We implemented the formula for both the expected values and the Log-Likelihood values as described by (Rayson and Garside, 2000).

Manual inspection of the Log-Likelihood figures confirmed our hypothesis that more domain-specific terms in our corpus were assigned high LL-values. We experimentally defined the threshold for Log-Likelihood values corresponding to distinctive terms on our development corpus. Example (4) shows some translation pairs which are filtered out by applying the LL threshold.

(4) Fr: cependant – En: however – It: tuttavia – Du: echter
Fr: choix – En: choice – It: scelta – Du: keuze
Fr: continuer – En: continue – It: continuare – Du: verdergaan
Fr: cadre – En: frame – It: cornice – Du: frame (erroneous filtering)
Fr: allégement – En: lightening – It: alleggerire – Du: verlichten (erroneous filtering)

4.1.2 Mutual Expectation Measure

The Mutual Expectation measure as described by Dias and Kaalep (2003) is used to measure the degree of cohesiveness between words in a text. This way, candidate multiword terms whose components do not occur together more often than expected by chance get filtered out. In a first step, we have calculated all n-gram frequencies (up to 8-grams) for our four automotive corpora and then used these frequencies to derive the Normalised

[1] http://catalog.elra.info/product_info.php?products_id=438

Expectation (NE) values for all multiword entries, as specified by the formula of Dias and Kaalep:

$$NE = \frac{prob(n-gram)}{\frac{1}{n}\sum prob(n-1-grams)} \quad (3)$$

The Normalised Expectation value expresses the cost, in terms of cohesiveness, of the possible loss of one word in an n-gram. The higher the frequency of the n-1-grams, the smaller the NE, and the smaller the chance that it is a valid multiword expression. The final Mutual Expectation (ME) value is then obtained by multiplying the NE values by the n-gram frequency. This way, the Mutual Expectation between n words in a multiword expression is based on the Normalised Expectation and the relative frequency of the n-gram in the corpus.

We calculated Mutual Expectation values for all candidate multiword term pairs and filtered out incomplete or erroneous terms having ME values below an experimentally set threshold (being below 0.005 for both source and target multiword or below 0.0002 for one of the two multiwords in the translation pair). The following incomplete candidate terms in example (5) were filtered out by applying the ME filter:

(5) Fr: fermeture embout - En: end closing - It: chiusura terminale - Du: afsluiting deel
(should be: Fr: fermeture embout de brancard - En: chassis member end closing panel - It: chiusura terminale del longherone - Du: afsluiting voorste deel van langsbalk)

4.2 Evaluation

The terminology extraction module was tested on all sentences from the three test corpora. The output was manually labeled and the annotators were asked to judge both the *translational quality* of the entry (both languages should refer to the same referential unit) as well as the *relevance* of the term in an automotive context. Three labels were used: OK (valid entry), NOK (not a valid entry) and MAYBE (in case the annotator was not sure about the relevance of the term).

First, the impact of the statistical filtering was measured on the bilingual term extraction. Secondly, we compared the output of our system with the output of a commercial bilingual terminology extraction module and with the output of a set of standard monolingual term extraction modules.

Since the annotators labeled system output, the reported scores all refer to precision scores. In future work, we will develop a gold standard corpus which will enable us to also calculate recall scores.

4.2.1 Impact of filtering

Table 5 shows the difference in performance for both single and multiword terms with and without filtering. Single-word filtering seems to have a bigger impact on the results than multiword filtering. This can be explained by the fact that our candidate multiword terms are generated from anchor chunks (chunks aligned with a very high precision) that already answer to strict syntactical constraints. The annotators also mentioned the difficulty of judging the relevance of single word terms for the automotive domain (no clear distinction between technical and common vocabulary).

	Not Filtered			Filtered		
	OK	NOK	MAY	OK	NOK	MAY
FR-EN						
Sing_w	82%	17%	1%	86.5%	12%	1.5%
Mult_w	81%	16.5%	2.5%	83%	14.5%	2.5%
FR-IT						
Sing_w	80.5%	19%	0.5%	84.5%	15%	0.5%
Mult_w	69%	30%	1.0%	72%	27%	1.0%
FR-DU						
Sing_w	72%	25%	3%	75%	22%	3%
Mult_w	83%	15%	2%	84%	14%	2%

Table 5: Impact of statistical filters on Single and Multiword terminology extraction

4.2.2 Comparison with bilingual terminology extraction

We compared the three filtered bilingual lexicons (French versus English-Italian-Dutch) with the output of a commercial state-of-the-art terminology extraction program SDL MultiTerm Extract[2]. MultiTerm is a statistically based system that first generates a list of candidate terms in the source language (French in our case) and then looks for translations of these terms in the target language. We ran MultiTerm with its default settings (default noise-silence threshold, default stopword list, etc.) on a large portion of our parallel corpus that also contains all test sentences[3]. We ran our system (where term extraction happens on a sentence per sentence basis) on the three test sets.

[2] www.translationzone.com/en/products/sdlmultitermextract

[3] 70,000 sentences seemed to be the maximum size of the corpus that could be easily processed within MultiTerm Extract.

Table 6 shows that even after applying statistical filters, our term extraction module retains a much higher number of candidate terms than MultiTerm.

	# Extracted terms	# Terms after filtering	MultiTerm
FR-EN	4052	3386	1831
FR-IT	4381	3601	1704
FR-DU	3285	2662	1637

Table 6: Number of terms before and after applying Log-Likelihood and ME filters

Table 7 lists the results of both systems and shows the differences in performance for single and multiword terms. Following observations can be made:

- The performance of both systems is comparable for the extraction of single word terms, but our system clearly outperforms MultiTerm when it comes to the extraction of more complex multiword terms.

- Although the alignment results for French-Italian were very good, we do not achieve comparable results for Italian multiword extraction. This can be due to the fact that the syntactic structure is very similar in both languages. As a result, smaller syntactic chunks are linked. However one can argue that, just because of the syntactic resemblance of both languages, the need for complex multiword terms is less prominent in closely related languages as translators can just paste smaller noun phrases together in the same order in both languages. If we take the following example for instance:

  déposer – l' embout – de brancard
  togliere – il terminale – del sotto-porta

  we can recompose the larger compound *l'embout de brancard* or *il terminale del sottoporta* by translating the smaller parts in the same order (*l'embout – il terminale* and *de brancard – del sottoporta*

- Despite the worse alignment results for Dutch, we achieve good accuracy results on the multiword term extraction. Part of that can be explained by the fact that French and Dutch use a different compounding strategy: whereas French compounds are created by concatenating prepositional phrases, Dutch usually tends to concatenate noun phrases (even without inserting spaces between the different compound parts). This way we can extract larger Dutch chunks that correspond to several French chunks, for instance:

  Fr: feu régulateur – de pression carburant.
  Du: brandstofdrukregelaar.

	ANCHOR CHUNK APPROACH			MULTITERM		
	OK	NOK	MAY	OK	NOK	MAY
FR-EN						
Sing_w	86.5%	12%	1.5%	77%	21%	2%
Mult_w	83%	14.5%	2.5%	47%	51%	2%
Total	84.5%	13.5%	2 %	64%	34%	2%
FR-IT						
Sing_w	84.5%	15%	0.5%	85%	14%	1%
Mult_w	72%	27%	1.0%	65%	34%	1%
Total	77.5%	22%	1%	76.5%	22.5%	1%
FR-DU						
Sing_w	75%	22%	3%	64.5%	33%	2.5%
Mult_w	84%	14%	2%	49.5%	49.5%	1%
Total	79.5%	20%	2.5%	58%	40%	2%

Table 7: Precision figures for our term extraction system and for SDL MultiTerm Extract

4.2.3 Comparison with monolingual terminology extraction

In order to have insights in the performance of our terminology extraction module, without considering the validity of the bilingual terminology pairs, we contrasted our extracted English terms with state-of-the art monolingual terminology systems. As we want to include both single words and multiword terms in our technical automotive lexicon, we only considered ATR systems which extract both categories. We used the implementation for these systems from (Zhang et al., 2008) which is freely available at[1].

We compared our system against 5 other ATR systems:

1. Baseline system (Simple Term Frequency)

2. Weirdness algorithm (Ahmad et al., 2007) which compares term frequencies in the target and reference corpora

3. C-value (Frantzi and Ananiadou, 1999) which uses term frequencies as well as unit-hood filters (to measure the collocation strength of units)

[1] http://www.dcs.shef.ac.uk/~ziqizhang/resources/tools/

4. Glossex (Kozakov et al., 2004) which uses term frequency information from both the target and reference corpora and compares term frequencies with frequencies of the multiword components

5. TermExtractor (Sclano and Velardi, 2007) which is comparable to Glossex but introduces the "domain consensus" which "simulates the consensus that a term must gain in a community before being considered a relevant domain term"

For all of the above algorithms, the input automotive corpus is PoS tagged and linguistic filters (selecting nouns and noun phrases) are applied to extract candidate terms. In a second step, stopwords are removed and the same set of extracted candidate terms (1105 single words and 1341 multiwords) is ranked differently by each algorithm. To compare the performance of the ranking algorithms, we selected the top terms (300 single and multiword terms) produced by all algorithms and compared these with our top candidate terms that are ranked by descending Log-likelihood (calculated on the BNC corpus) and Mutual Expectation values. Our filtered list of unique English automotive terms contains 1279 single words and 1879 multiwords in total. About 10% of the terms do not overlap between the two term lists. All candidate terms have been manually labeled by linguists. Table 8 shows the results of this comparison.

	Single Word terms			Multiword terms		
	OK	NOK	MAY	OK	NOK	MAY
Baseline	80%	19.5%	0.5%	84.5%	14.5%	1%
Weirdness	95.5%	3.5%	1%	96%	2.5%	1.5%
C-value	80%	19.5%	0.5%	94%	5%	1%
Glossex	94.5%	4.5%	1%	85.5%	14%	0.5%
TermExtr.	85%	15%	0%	79%	20%	1%
AC approach	85.5%	14.5%	0%	90%	8%	2%

Table 8: Results for monolingual Term Extraction on the English part of the automotive corpus

Although our term extraction module has been tailored towards bilingual term extraction, the results look competitive to monolingual state-of-the-art ATR systems. If we compare these results with our bilingual term extraction results, we can observe that we gain more in performance for multiwords than for single words, which might mean that the filtering and ranking based on the Mutual Expectation works better than the Log-Likelihood ranking.

An error analysis of the results leads to the following insights:

- All systems suffer from partial retrieval of complex multiwords (e.g. ATR *management ecu* instead of *engine management ecu*, AC approach *chassis leg end piece closure* instead of *chassis leg end piece closure panel*).

- We manage to extract nice sets of multiwords that can be associated with a given concept, which could be nice for automatic ontology population (e.g. AC approach *gearbox casing, gearbox casing earth, gearbox casing earth cable, gearbox control, gearbox control cables, gearbox cover, gearbox ecu, gearbox ecu initialisation procedure, gearbox fixing, gearbox lower fixings, gearbox oil, gearbox oil cooler protective plug*).

- Sometimes smaller compounds are not extracted because they belong to the same syntactic chunk (E.g we extract *passenger compartment assembly, passenger compartment safety, passenger compartment side panel*, etc. but not *passenger compartment* as such).

5 Conclusions and further work

We presented a bilingual terminology extraction module that starts from sub-sentential alignments in parallel corpora and applied it on three different parallel corpora that are part of the same automotive corpus. Comparisons with standard terminology extraction programs show an improvement of up to 20% for bilingual terminology extraction and competitive results (85% to 90% accuracy) for monolingual terminology extraction. In the near future we want to experiment with other filtering techniques, especially to measure the domain distinctiveness of terms and work on a gold standard for measuring recall next to accuracy. We will also investigate our approach on languages which are more distant from each other (e.g. French – Swedish).

Acknowledgments

We would like to thank PSA Peugeot Citroën for funding this project.

References

K. Ahmad, L. Gillam, and L. Tostevin. 2007. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and rerieval (wilder). In *Proceedings of the Eight Text REtrieval Conference (TREC-8)*.

S. Ananiadou. 1994. A methodology for automatic term recognition. In *Proceedings of the 15th conference on computational linguistics*, pages 1034–1038.

R.H. Baayen, R. Piepenbrock, and H. van Rijn. 1993. The celex lexical database on cd-rom.

I. Dagan and K. Church. 1994. Termight: identifying and translating technical terminology. In *Proceedings of Applied Language Processing*, pages 34–40.

B. Daille. 1995. Study and implementation of combined techniques for automatic extraction of terminology. In J. Klavans and P. Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. MIT Press, Cambridge, Massachusetts; London, England.

G. Dias and H. Kaalep. 2003. Automatic extraction of multiword units for estonian: Phrasal verbs. *Languages in Development*, 41:81–91.

K.T. Frantzi and S. Ananiadou. 1999. the c-value/nc-value domain independent method for multiword term extraction. *journal of Natural Language Processing*, 6(3):145–180.

K. Kageura and B. Umino. 1996. Methods of automatic term recognition: a review. *Terminology*, 3(2):259–289.

L. Kozakov, Y. Park, T.-H Fin, Y. Drissi, Y.N. Doganata, and T. Confino. 2004. Glossary extraction and knowledge in large organisations via semantic web technologies. In *Proceedings of the 6th International Semantic Web Conference and he 2nd Asian Semantic Web Conference (Se-mantic Web Challenge Track)*.

J. Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*.

L. Macken and W. Daelemans. 2009. Aligning linguistically motivated phrases. In van Halteren H. Verberne, S. and P.-A. Coppen, editors, *Selected Papers from the 18th Computational Linguistics in the Netherlands Meeting*, pages 37–52, Nijmegen, The Netherlands.

L. Macken, E. Lefever, and V. Hoste. 2008. Linguistically-based sub-sentential alignment for terminology extraction from a bilingual automotive corpus. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 529–536, Manchester, United Kingdom.

R. C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, Machine Translation: from research to real users, pages 135–244, Tiburon, California.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, pages 1–6.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

F. Sclano and P. Velardi. 2007. Termextractor: a web application to learn the shared terminology of emergent web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*.

A. Van den Bosch, G.J. Busser, W. Daelemans, and S. Canisius. 2007. An efficient memory-based morphosyntactic tagger and parser for dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114, Leuven, Belgium.

V. Vandeghinste. 2008. *A Hybrid Modular Machine Translation System. LoRe-MT: Low Resources Machine Translation.* Ph.D. thesis, Centre for Computational Linguistics, KULeuven.

Z. Zhang, J. Iria, C. Brewster, and F. Ciravegna. 2008. A comparative evaluation of term recognition algorithms. In *Proceedings of the sixth international conference of Language Resources and Evaluation (LREC 2008)*.