From Statistical Term Extraction to Hybrid Machine Translation

Petra Wolf, Ulrike Bernardi
Lucy Software and Services
Neumarkter Str. 81
81673 Munich, Germany
petra.wolf,ulrike.bernardi@lucysoftware.com

Christian Federmann, Sabine Hunsicker
DFKI
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
sabine.hunsicker,christian.federmann@dfki.de

Abstract

This study presents a new hybrid approach for translation equivalent selection within a transfer-based machine translation system using an intertwined net of traditional linguistic methods together with statistical techniques. Detailed evaluation reveals that the translation quality can be improved substantially in this way.

1 Introduction

A promising integration point for statistical techniques into Rule-Based Machine Translation (RBMT) systems is the transfer phase. A key problem here is to deal with non-deterministic rules and preferences in order to disambiguate and select the most natural expressions in the target language (cf. (Eisele et al., 2008b; Thurmair, 2009)). We performed studies on a phrase-table-driven transfer approach based on an RBMT system (for details on the RBMT system see (Alonso and Thurmair, 2003)): A prototype was built accessing statistically generated bilingual phrase tables at runtime in addition to system lexicons and grammars. This hybrid prototype showed indeed better lexical selection, improved coverage and even sometimes enhanced syntax, but well-known issues in morpho-syntax and a very low hit rate of the phrase table module remained as limiting factors. Thus although the data which were accessed and preferred are obviously the desired ones, the first challenge was the lack of deeper linguistic knowledge within the data while the second challenge was the small F-measure of the data itself.

This led us to a deeper intertwined hybrid extension: The **LiSTEX** approach (Hybrid Transfer by multi-layered **Li**nguistically augmented **S**tatistical **T**erminology **EX**traction). We decided to impose strict knowledge-enhanced requirements for the statistical term extraction which consequently was augmented by several layers of linguistic data refinement and automatic feature attribution. **LiSTEX** is distinct from other term extractors as it has layers that access already during the early term defining extraction phase the RBMT system components for linguistic filtering and later production of knowledge-augmented output. In contrast to the phrase-table approach, the F-measure is high which avoids evident deteriorations after integration into the MT system.

The remainder of this paper is organized as follows: Section 2 gives an overview of related work. Details on the terminology extraction are then provided in Section 3 followed by a description of the results of terminology and translation quality evaluations in section 4. Section 5 summarizes the key findings and outlines open issues for future work.

2 Related work

There have been several studies of hybrid machine translation approaches to overcome the drawbacks of rule-based and statistical MT alone by a combined approach, starting from RBMT as well as starting from SMT. Evaluations of SMT vs. RBMT systems revealed that one of the weak points of RBMT systems is the lexical selection in transfer (cf. (Thurmair, 2009; Chen et al., 2009). Since RBMT systems tend to suffer from insufficient and too deterministic lexical coverage and choice (Eisele et al., 2008b), this study concentrates on automatic enlargements of RBMT lexicons and enhanced transfer-generation operations, while taking into account the peculiarities of a specific RBMT system and statistical techniques.

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 225–232
Leuven, Belgium, May 2011

The considerable potential of statistical term extraction combined with RBMT has been evaluated by other researchers, such as (Thurmair, 2003; Dugast et al., 2009). Here we go one step further by already integrating some RBMT techniques into the very early stages of the statistical term extraction process. This helps to avoid the well-known problems found in term guessing and identification (Heid, 1999). Additional linguistic layers assure that the extracted terms and phrases are tailored and augmented for the RBMT system in question.

3 Intelligent Terminology Extraction

The underlying term extraction tool extracts term pairs by means of statistical algorithms from existing translation memories or bilingual corpora (Eisele et al., 2008a). As a bilingual corpus, the Europarl corpus (Koehn, 2005) was used for development. As a test set, we choose the ACL WMT 2008 test set which contains the Q4/2000 portion of the EuroParl data (2000-10 to 2000-12).

The initial design study on peculiarities of a specific statistically based term extraction that satisfies the requirements of an RBMT system dealt with issues like term definition in a strict linguistic and system-related sense, term identification as single words vs. multiword expressions, base form reduction and application of linguistic category patterns for identification or elimination of noise at the end of the first multi-layered extraction phase. Also the term recognition needed to be defined in relation with the linguistically based target system, i.e. comparison of the extraction result with the target system lexicon in order to identify known vs. unknown terms. Thus this painstaking specification research provided the basis for the extended implementation of the LiSTEX term identification, knowledge-enhanced acquisition and augmentation modules.

For LiSTEX, we concentrated on German-English and Spanish-English. The remaining subsections explain details of the intelligent term extraction techniques and present the term extraction workflow as a whole (see Figure 1).

3.1 Term Identification with Initial Meta Knowledge Acquisition and Organization

The objective is to have the semantic translation equivalents decided by the statistical technique. The requirements, however, on the linguistic contents of this terminology to be extracted and imported are determined by the fine-grained data needed by an RBMT system. The first major modification of the statistical extraction has to fulfill the requirement to extract only well defined linguistic terms. For this reason, the extraction toolkit has been extended in order to access and use the RBMT analysis, transfer and generation trees to find interesting linguistic phrases, i.e. terms. In addition, it checks whether the translation of a specific term in the transfer tree, be it a single or a multiword, differs from the reference translation in the corpus. In this way, the term extraction tool delivers only terms which receive a linguistic identification and which are translated differently with the conventional RBMT system than in the bilingual memory.

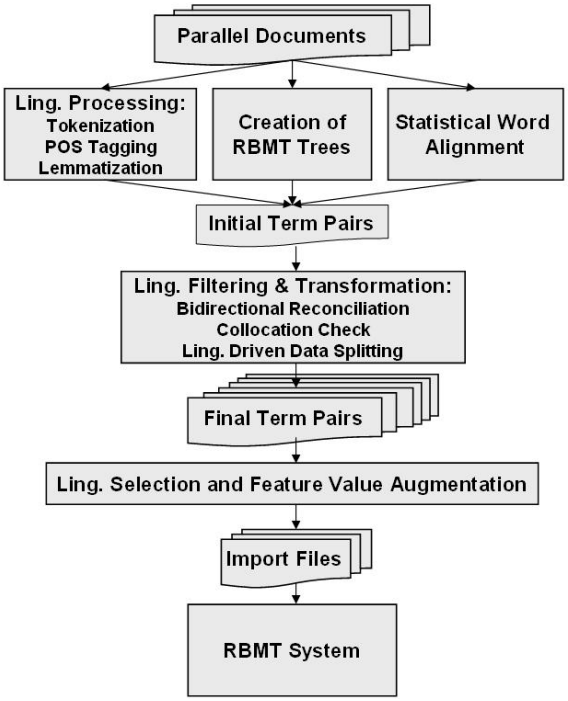

Figure 1: Term Extraction Workflow

The extracted terminology has to contain the following minimum information: Canonical forms and categories of the source and target terms, domain area, frequency and, in case of multiwords, the internal structure of the expression. Finally, the context has to be passed through: One example source language sentence in which the term occurs in the corpus with the corresponding translation equivalent sentence, useful for later manual inspection.

For achieving this first layer of terminology acquisition, the source and reference text are tok-

enized, tagged with part of speech information and the tokens are lemmatized. Source and reference texts are aligned word-by-word, the source text is translated by the RBMT Engine and analysis, transfer and generation trees are created. These trees are aligned to the words in the source and reference text and all phrases which the RBMT system translated differently than the reference translation are selected as potential term candidates.

Since these phrases still include the original surface forms, the canonical forms are built now, e.g. adjectives are inflected properly to match the head noun. Thus the terms receive the proper dictionary format. The categories for the entire terms are derived from the part of speech sequences. The list of terms is alphabetically sorted and non-frequent terms are filtered out.

Up to this point in the processing chain, the extraction process is done for one language direction only. Now the other language direction is generated and the intersection of both is created to avoid wrong term pairs caused by alignment problems.

The next step is the collocation check to reduce the amount of false entries. Frequencies for all term pairs (single words as well as multiword expressions) are extracted. We check the single word entries and extract all the multiword expressions containing this word and their corresponding frequencies. If a word is part of a multiword, a certain threshold calculation referring to the frequencies of the single word and its corresponding multiwords limits the acquisition algorithm of single words. The higher this threshold is set, the more single terms will be rejected. In this way, we can avoid that the proper name *Tour de France* in the text will lead to two extracted terms: *Tour de France* and *France*, since *France* only appears within that multiword and not on its own and is therefore not a correct translation into German.

Once these refined extraction layers have been completed, there are final actions to be performed in order to generate terminology which is linguistically more precise to be handled accurately during the subsequent cleanup and feature augmentation preprocessing. Thus the next step performs an additional **linguistically driven differentiation process** important for quality growth.

- Entries where the **lemmatizer could not find the correct lemma** and just guesses it, are stored in separate lists.
- **Entries containing special characters**, such as punctuation marks or integers, are filtered into separate lists. Since these entries tend to be wrongly aligned and extracted, the sublists have to be inspected and only useful terms will be imported.
- All entries showing a **category change** (i.e. source and target categories are not identical) are excluded from import since they are likely to be wrong. Sometimes they are caused by errors during part of speech tagging, sometimes by alignment errors.
- **Quality Splitting** procedures which allow further predictions as to the quality of the extracted terms. A German multiword term for example which results in an English single target word is a likely non-valid term, mostly due to alignment errors:
  1. Single words on source and target side
  2. Single words on source, multiword expressions on target side
  3. Multiword expressions on source side, single words on the target side
  4. Multiword expressions for source and target language
- Additional generation of **part of speech sublists** for nouns, verbs, adjectives and adverbs respectively will facilitate the import preprocessing since the linguistic splitting of the output files allows for a fast and precise quality assessment of intermediate modules. Moreover, the further automatic linguistic feature augmentation of the extracted terms will be category-dependent.

3.2 RBMT-Targeted Quality Precision Phase

After completion of the LiSTEX acquisition phase, semi-automatic inspection of the filtered and split terms revealed several problems mainly due to:

- Wrong derivation of verbalized nouns: *accuse, belong* instead of *accused, belonging*.
- Missing noun heads: *monetary* instead of *monetary authority*, *automobile* instead of *automobile industry*.
- Weak adjectival nouns which would be wrongly analyzed as plural: *Angeklagte im Strafverfahren* instead of the correct masculine canonical form *Angeklagter im Strafverfahren*.
- All kinds of wrong derivation of adjectival endings: *medizinisch Versorgung,*

gesundheitsbezogenere Angabe instead of *medizinische Versorgung (medical care), gesundheitsbezogene Angabe (health-related claim)*.

- Non-basic forms or non-existent forms like *Basler Ausschusses*, *Brüsselrer Flughafen* instead of *Basler Ausschuss*, *Brüsseler Flughafen*.
- Wrong Plurals: *Genitalverstümmelungen bei Frauen → female genital mutilation*.
- Varying capitalization, thus variants' multiplication: *Eu Presidency, Eu presidency, EU Presidency*.
- Wrong category assignment: noun instead of adverb for the expressions *case by case*.
- US vs. UK spelling
- Wrong translation equivalents like: *Kandidatur der Türkei → Turkey 's*, *Grundgesetz von Hongkong → of of*.

The majority of these false terms were due to lemmatizer and derivational errors in the linguistic submodules of the term acquisition. To a great extent these wrong terms are corrected in the following **precision phase**: Wrong or missing endings of adjective specifiers were rectified, wrong superlative adjective specifier compounds and incorrect derivation of irregular adjective specifiers were corrected. Senseless terms, such as *of of*, and wrongly categorized entries, such as *belong, deepen, Foreign* as nouns are deleted. Single word nouns in plural form are evaluated to avoid getting duplicates to already existing singular entries.

Also most of the wrong translation equivalents which followed a certain pattern or syntax could be deleted. Of course, a certain number of wrong equivalents, especially the pure semantic ones, can only be detected and deleted manually. Since we did only automatic cleanup, these errors will appear later as specific translation errors, but they will only play a minor role (cf. Section 4).

A major role in the quality refinement as a whole is played by the **selection process** which benefits now from the former differentiation process. Term pairs with very improbable category changes (i.e. if the source term is a noun and the target term an adverb) will be completely deleted. Also for example all acquired term data containing a German multiword source with an English single target word will be discarded and thus cannot undergo the further transaction of automatic augmentation (example: *europäischer Rat im Dezember → December*). In addition, most of the extracted adjectives, adverbs or verbs after intersection with the RBMT lexicons are wrong and will not be imported.

Approx. 2% of German-English entries and almost 4% of the English-Spanish data were corrected by these automatic cleanup procedures. About 10% of the originally extracted entries are deleted for German-English and around 25% for English-Spanish (cf. Table 2 for details on the number of extracted terms and their error rate before and after semi-automatic inspection, automatic cleanup and selection).

3.3 Linguistic Feature-Value-Pair Augmentation

After completion of the refined term extraction, the next step involves deep feature and value assignment. Targeted category- and content-specific algorithms for the respective linguistically classified terms could be developed and applied: The import preprocessing parses and augments the data structures by generating the obligatory linguistic feature-value pairs for the RBMT system. Furthermore, it creates the monolingual and bilingual information for the three system lexicons (two monolingual and one bilingual lexicon) and finally writes the entries.

This is performed by the subsequent module containing the **Input Parser** and the **Defaulter**: The lexicographic information needed for a complete RBMT entry can be automatically created from individual parts of the entry so far generated. Even information from one entry can be used to complete other entries so that available information in already existing RBMT lexicons can be accessed and used for calculation of new entries:

Single Word Defaulting When the user imports a new monolingual entry tagged with noun, verb, adjective, adverb, all the obligatory feature values (e.g. for nouns: allomorph, declension class value, linguistic gender, kind of noun (e.g. proper nouns, countable nouns), natural sex, semantic type of noun, e.g. abstract nouns, temporal units and the like) can automatically be inserted by LiSTEX. These defaults represent the best *guess* the system can make.

Multiword Defaulting Multiwords consist of heads and variable parts. The input parser can automatically parse and recognize their internal structure and then creates monolingual entries for

heads and variable parts and defaults the obligatory values.

Defaulting Monolingual Entries from Transfer Entries Transfer entries are not usable in the RBMT system if there are no corresponding monolingual entries for analysis and generation. Therefore, after processing the transfer entries, the corresponding mono entries are automatically generated.

Deriving There are features that logically follow from already defined features. For example, gender and number features are derived from the class value. These features are not generated in the same way as the defaulted features are. Defaulting, after all, is only a kind of intelligent guess. Derived features, on the other hand, are calculated and by definition are correct, provided they are deduced from correct feature-values.

After completing this input parsing and defaulting step, the so far extracted translation equivalent term *Beitrittsverhandlung → accession negotiation* with the annotation noun is augmented in this phase and receives the following automatically generated structures and feature-value-pairs: Defaulted monolingual and bilingual entries:

- (”Beitrittsverhandlung” NST ALO ”Beitrittsverhandlung” ARGS ((N1 (PREP ”mit” ”über”) (CA A)) (N0 (PREP0 ”über”) (FCP TH) (INT T))) CL (P-EN S-0) GD (F) KN MS-CNT LINK S SX (N) TYN (PRO) !AUTHOR ”TermExtract” !OWNER ”sys” !DATE 1296734190)
- (”accession negotiation” NST ALO ”accession negotiation” KN CNT TYN (ABS PRO) MW-TYPE STRING-NST MW-BODY ((STRING ”accession”) (HEAD)) !AUTHOR ”TermExtract” !OWNER ”sys” !DATE 1296734198 MW-HEAD-CAN ”negotiation” MW-HEAD-CAT NST)
- (”Beitrittsverhandlung” NST ”accession negotiation” NST PRF 10000 TAG (POL) !DATE 1296734170 !OWNER ”sys” !AUTHOR ”TermExtract”)

As one can see from both monolingual entries, the defaulting mechanism before making its guesses opens the already existing lexicons and accesses - if available - already coded information, here from the entries *Verhandlung* and *negotiation* and propagates the identified feature-value pairs on the corresponding new entries.

4 Evaluations

4.1 Extraction and Import Quantification

For the evaluation, we extracted phrase lists from the EuroParl corpus (Koehn, 2005): For German-English, the Europarl Corpus contains 1,259,571 lines and for Spanish-English 1,253,026 lines. Table 1 shows a large number of monolingual entries compared to the relatively small number of bilingual entries. This can be explained by the fact that the monolingual lists still contain many entries which only appear once in the corpus. Many of these entries are faulty, so they are discarded early on. For example, for German→ English 365,243 of the 441,425 terms have a frequency of 1.

Translation Direction	Monoling. List	Biling. List
German-English	441,425 / 508,592	45,857
Spanish-English	406,296 / 294,069	35,088

Table 1: Extracted Terms

The bilingual lists form the basis for removing unwanted single terms resulting from multiwords. In the end, we got 30,803 terms for German-English with an error rate of about 14% (cf. Table 2). After the semi-automatic correction, the number of entries is reduced to 27,054 terms with an error rate of about 8%. We evaluated the error rate by randomly selecting 5% of the whole data set and interpolated the results to the complete data set. For English-Spanish, the results are slightly better with an error reduction from 18.25% to 3.8%.

The corrected and augmented terms are finally imported into the RBMT lexicons. For German-English, out of the 27,054 terms, around 26,000 entries could be imported without any further action. 931 term pairs for German → English respectively 720 for English → German result in conflicts during the import. Namely, conflicting transfer entries are term pairs which already exist in the RBMT lexicon, but with additional information, such as additional tests or transformations performed during transfer. Therefore, the new entries have been added to the lexicon in addition to the already existing entries in order not to loose information. Furthermore, approx. 1,500 transfer entries were modified during the import, since these are term pairs which are identical to existing RBMT lexicon entries, but just differ in the subject area information. These entries are merged with the existing lexicon entries during import. Only for about half of the terms new monolingual English or German entries were created, since the other half is already available in the lexicon. On one hand, this is due to the fact that the RBMT lexicons are already quite big and contain a great variety of terminology. On the other hand, the Europarl cor-

	English ↔ Spanish		English ↔ German	
	Extracted Terms	Terms after Correction	Extracted Terms	Terms after Correction
Number of Terms	24,519	18,546	30,803	27,054
Error Rate	18.25%	3.8%	14.25%	7.71%

Table 2: Error Rates for Extracted and Corrected Term Pairs

pus does not contain very specific terminology, but more general terms which are already covered by the RBMT lexicons.

The results for the English-Spanish import are very similar to the ones for German-English: Out of the 18,546 terms, around 18,000 entries could be imported without any further action, whereas approx. 500 term pairs resulted in conflicts during the import and have been added as additional entries. Nearly 2,000 entries have been modified and merged with existing entries just differing in the subject area information. Here again for only half of the terms new monolingual English or Spanish entries were created, since the other entries already exist in the RBMT lexicon.

4.2 Translation-Related Evaluation

The translation evaluation after the LiSTEX import revealed a high number of differently translated sentences: All four directions showed more than 95% differently translated sentences (cf. Table 4). And even within these sentences we found not only one, but several differences.

The BLEU and NIST scores did not show any improvements comparing LiSTEX to the baseline system (cf. Table 3). But since BLEU's correlation with human judgments has already been questioned (Callison-Burch et al., 2006), we performed a manual evaluation which revealed clear improvements for all language directions.

For this evaluation, we adapted the Appraise evaluation tool (Federmann, 2010) so that the interface shows the source sentence, reference translation and the two anonymised translation candidates by the RBMT system. Anonymising the order will eliminate potential bias by the human annotators.

From these different translations one third up to half of them are of equal quality as before. This portion is mainly due to new alternatives which may vary without changing the quality of the new output. This outcome is due to the fact that the RBMT lexicons are of quite broad coverage and that the Europarl corpus covers mostly general vocabulary.

In all language directions, the translation quality evaluation revealed clear improvements, ranging from approx. 38% better translations in Spanish → English to more than 50% improved translations in German → English. However, these improvements are offset by approx. 10% to 20% worsened translations. In the following, we evaluate in detail all the translation differences in order to find out the reasons for the big variety between the various language directions on one hand and for the deteriorations caused by LiSTEX on the other hand to develop mechanisms to compensate them.

Improvements The translation evaluation revealed several improvements: More appropriate terminology is used. For example, the term *prostitución infantil* is now correctly translated as *child prostitution*. Before the expression was compositionally translated as *infantile prostitution* which is wrong.

Even the recognition of the whole sentence structure can occur to be substantially improved with the additional terminology, since the compositional analysis of complex multiword structures is no longer necessary, if these multiwords are now in the lexicon.

Multiword Parts Added and/or Lost Sometimes during extraction a part of a multiword has been lost or in contrary was added. If not corrected during the precision phase, the result is a wrong entry which in consequence produces false translations. For example, *Zeitraum* is wrongly translated as *five-year period* instead of *period*.

Wrong Translation Equivalents There are translations that have underlying inadequate term equivalents in a given context, i.e. *November* is translated into *October* since the term imported was derived from the human translation mistake made in the original corpus. Sometimes it is a specific idiomatic usage in the corpus which then after extraction does not reflect the default usage.

	Baseline		LiSTEX	
Translation Direction	NIST	BLEU	NIST	BLEU
German → English	5.5582	0.1632	5.2430	0.1491
English → German	4.3616	0.1078	4.2718	0.1060
Spanish → English	5.8776	0.1953	5.6414	0.1830
English → Spanish	6.1375	0.2085	5.9086	0.1978

Table 3: BLEU and NIST Scores

	English ↔ Spanish		English ↔ German	
	English → Spanish	Spanish → English	English → German	German → English
Translated TUs	2,000	2,000	2,000	2,000
Different Translations	95.05%	96.05%	95.45%	96.20%
Evaluated Differences	287	398	970	1,261
Better	47.74%	38.44%	50.82%	57.49%
Equal	31.71%	46.73%	41.86%	30.29%
Worse	20.56%	14.82%	7.32%	12.21%

Table 4: Translation Quality Evaluation of English-Spanish and English-German

Multiword Expressions Multiword expressions in contrast to the fixed expressions maintain a large variability as to their morpho-syntactic behaviour. As we now import a huge number of multiwords and even collocations, the RBMT multiword treatment is challenged to a considerable degree. Especially coordination and PP-attachment procedures are heavily affected. Multiwords are normally fixed in their structure so that it is not possible to add additional modifiers to the head or variable parts. For collocations, this is slightly different as they interact more freely. If you import a term pair, such as *europäischer Rat in Nizza - Nice Council* and now translate the phrase *des Europäischen Rates in Nizza und in Biarritz*, the analysis can no longer attach *und in Biarritz* to the head *Rat* and therefore the analysis fails. Here more intelligent mechanisms have to be developed to process collocations and still allow for further modifying them.

Alternatives from TermExtraction Alternatives originating from the terminology extraction output have to be evaluated more specifically. At the moment, we reduced the import to the 5 best alternatives for a given source term according to their frequencies. These terms are then alphabetically ordered in the translation process. First evaluations revealed that the number of alternatives should be restricted even more and that apart from the scoring algorithm additional measures like mappings, default scores, frequency ordering are necessary.

Wrong Capitalization of English Terms The capitalization of the extracted English terms is inconsistent and thus produces capitalized and non-capitalized translation alternatives.

Missing Subcategorization Frames Since we added the extracted and augmented terminology, already existing entries which only have transformations and no tests are no longer preferred in transfer and thus imported entries with no subcategorization information are selected: nouns like *Bemerkung zu jdn./etw.* → *remark on sb./sth.* have in the original RBMT lexicon a transformation which had mapped the original preposition into the appropriate target preposition.

Quantifying the worse translations, the following picture can be drawn as to the given deterioration types in descending order: The type **Alternatives** is by far the most frequently disturbing factor followed by the extraction-error of **Multiword Part Lost/Added**. The third item of the **Multiword** errors may cause syntactic deterioration and more phrasal analyses although the entry itself has been correct. **Capitalization** as fourth error factor can also cause syntactic problems, since these capitalized terms are expected to be proper names. Thus these entries get special default values. The error type **Wrong Translation Equivalents** appears only on the final position. It could be reduced to a quite insignificant amount, so for the future work we will rather concentrate on the first four items cited in this list.

5 Conclusion

There is no simple press-a-button-method for filling RBMT systems with noticeable quality gain in the end. Our research, however, could clearly attest the great potential of the multi-level hybrid approach of LiSTEX: On one side we have an extended intertwined acquisition and precision setup which benefits from statistical techniques in accessing the most appropriate and frequent translations of terms and phrases in large corpora while using at the same time deeper knowledge like tree structures from the target RBMT system and linguistics-based technologies. On the other side, LiSTEX features a complex system using precise linguistic deriving and defaulting techniques for automatic execution of the obligatory feature-value augmentation and linguistic entity generation before system integration takes places.

We could show that almost all intermediate steps up to the final integration can be automized to a high degree which indeed is a convincing performance. As we could see in the evaluation description there are specific technological hurdles to overcome. For further verification of LiSTEX there is now need for other underlying extraction corpora which are more domain-specific and have little intersection with the already available broad-coverage RBMT system. Lemmatizer improvements need to be performed to ease the task of the subsequent precision module.

Another issue is the large amount of imported multiword and phrase-like entries which are rather free collocations than fixed idioms. By LiSTEX we could overcome the morpho-syntactic issues still showing up in SMT-bound systems, but now the level of challenge has moved up to the issue of deeper syntactic inclusion of larger linguistic phrase units which stand above fixed expressions and idioms in the strict sense, but which precisely play the major role when targeting for non-deterministic natural expressions in the translation output. This challenge needs to be addressed and additional adaptations have to be implemented since the RBMT grammars could perform such a proper handling of syntactically free collocations.

Further research is necessary to refine the selection among the translation alternatives produced by LiSTEX concentrating on additional narrow-down-procedures selecting by frequency, default and mapping measures.

Acknowledgment

Part of this work was carried out within the EuroMatrix Plus project and has been funded by the European Union as FP7-ICT-2007-3-231720.

References

Alonso, J. and G. Thurmair. 2003. The Comprendium Translator system. In *Proceedings of the MT Summit IX (New Orleans)*.

Callison-Burch, C., M. Osborne, and P. Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proceedings of the 11th EACL (Trento)*, pages 249–256.

Chen, Y., M. Jellinghaus, A. Eisele, Yi Z., S. Hunsicker, S. Theison, Ch. Federmann, and H. Uszkoreit. 2009. Combining multi-engine translations with Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (Athens)*, pages 42–46.

Dugast, L., J. Senellart, and P. Koehn. 2009. Statistical post editing and dictionary extraction: Systran/Edinburgh submissions for ACL-WMT2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (Athens)*, pages 110–114.

Eisele, A., C. Federmann, H. Uszkoreit, H. Saint-Amand, M. Kay, M. Jellinghaus, S. Hunsicker, T. Herrmann, and Y. Chen. 2008a. Hybrid architectures for multi-engine machine translation. In *Proceedings of Translating and the Computer 30*.

Eisele, A., C. Federmann, H. Uszkoreit, H. Saint-Amand, M. Kay, M. Jellinghaus, S. Hunsicker, T. Herrmann, and Y. Chen. 2008b. Hybrid machine translation architectures within and beyond the euromatrix project. In *Proceedings of European Association of Machine Translation (Hamburg)*, pages 27–34.

Federmann, C. 2010. Appraise: An open-source toolkit for manual phrase-based evaluation of translations. In *Proceedings of the Conference on International Language Resources and Evaluation (Valetta)*, pages 1731–1734.

Heid, U. 1999. Extracting terminologically relevant collocations from german technical texts. In *Proceedings of the International Congress on Terminology and Knowledge Engineering*, pages 241–255.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X (Phuket)*, pages 79–86.

Thurmair, G. 2003. Making term extraction tools usable. In *Proceedings of the Conference of Controlled Language Translation (Dublin)*, pages 170–179.

Thurmair, G. 2009. Comparing different architectures of hybrid machine translation systems. In *Proceedings of the MT Summit XII (Ottawa)*, pages 340–347.