ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora

Mārcis Pinnis[1], Radu Ion[2], Dan Ştefănescu[2], Fangzhong Su[3], Inguna Skadiņa[1], Andrejs Vasiļjevs[1], Bogdan Babych[3]

[1]Tilde, Vienības gatve 75a, Riga, Latvia
{marcis.pinnis,inguna.skadina,andrejs}@tilde.lv

[2]Research Institute for Artificial Intelligence, Romanian Academy
{radu,danstef}@racai.ro

[3]Centre for Translation Studies, University of Leeds
{f.su,b.babych}@leeds.ac.uk

Abstract

The lack of parallel corpora and linguistic resources for many languages and domains is one of the major obstacles for the further advancement of automated translation. A possible solution is to exploit comparable corpora (non-parallel bi- or multi-lingual text resources) which are much more widely available than parallel translation data. Our presented toolkit deals with parallel content extraction from comparable corpora. It consists of tools bundled in two workflows: (1) alignment of comparable documents and extraction of parallel sentences and (2) extraction and bilingual mapping of terms and named entities. The toolkit pairs similar bilingual comparable documents and extracts parallel sentences and bilingual terminological and named entity dictionaries from comparable corpora. This demonstration focuses on the English, Latvian, Lithuanian, and Romanian languages.

Introduction

In recent decades, data-driven approaches have significantly advanced the development of machine translation (MT). However, lack of sufficient bilingual linguistic resources for many languages and domains is still one of the major obstacles for further advancement of automated translation. At the same time, comparable corpora, i.e., non-parallel bi- or multilingual text resources such as daily news articles and large knowledge bases like Wikipedia, are much more widely available than parallel translation data.

While methods for the use of parallel corpora in machine translation are well studied (Koehn, 2010), similar techniques for comparable corpora have not been thoroughly worked out. Only the latest research has shown that language pairs and domains with little parallel data can benefit from the exploitation of comparable corpora (Munteanu and Marcu, 2005; Lu et al., 2010; Smith et al., 2010; Abdul-Rauf and Schwenk, 2009 and 2011).

In this paper we present the ACCURAT toolkit[1] - a collection of tools that are capable of analysing comparable corpora and extracting parallel data which can be used to improve the performance of statistical and rule/example-based MT systems.

Although the toolkit may be used for parallel data acquisition for open (broad) domain systems, it will be most beneficial for under-resourced languages or specific domains which are not covered by available parallel resources.

The ACCURAT toolkit produces:

- **comparable document pairs** with comparability scores, allowing to estimate the overall comparability of corpora;
- **parallel sentences** which can be used as additional parallel data sources for statistical translation model learning;

[1] http://www.accurat-project.eu/

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 91–96,
Jeju, Republic of Korea, 8-14 July 2012.

- **terminology dictionaries —** this type of data is expected to improve domain-dependent translation;
- **named entity dictionaries**.

The demonstration showcases two general use case scenarios defined in the toolkit: "parallel data mining from comparable corpora" and "named entity/terminology extraction and mapping from comparable corpora".

The next section provides a general overview of workflows followed by descriptions of methods and tools integrated in the workflows.

1 Overview of the Workflows

The toolkit's tools are integrated within two workflows (visualised in Figure 1).

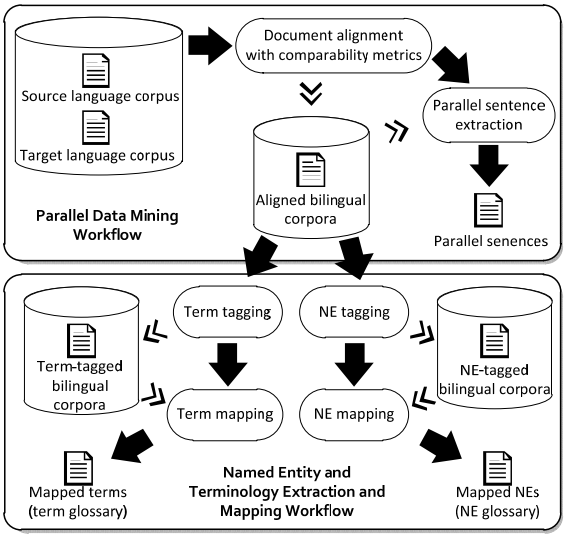

Figure 1. Workflows of the ACCURAT toolkit.

The **workflow for parallel data mining from comparable corpora** aligns comparable corpora in the document level (section 2.1). This step is crucial as the further steps are computationally intensive. To minimise search space, documents are aligned with possible candidates that are likely to contain parallel data. Then parallel sentence pairs are extracted from the aligned comparable corpora (section 2.2).

The **workflow for named entity** (NE) **and terminology extraction and mapping** from comparable corpora extracts data in a dictionary-like format. Providing a list of document pairs, the workflow tags NEs or terms in all documents using language specific taggers (named entity recognisers (NER) or term extractors) and performs multi-lingual NE (section 2.3) or term mapping (section 2.4), thereby producing bilingual NE or term dictionaries. The workflow also accepts pre-processed documents, thus skipping the tagging process.

Since all tools use command line interfaces, task automation and workflow specification can be done with simple console/terminal scripts. All tools can be run on the Windows operating system (some are also platform independent).

2 Tools and Methods

This section provides an overview of the main tools and methods in the toolkit. A full list of tools is described in ACCURAT D2.6. (2011).

2.1 Comparability Metrics

We define comparability by how useful a pair of documents is for parallel data extraction. The higher the comparability score, the more likely two documents contain more overlapping parallel data. The methods are developed to perform lightweight comparability estimation that minimises search space of relatively large corpora (e.g., 10,000 documents in each language). There are two comparability metric tools in the toolkit: a translation based and a dictionary based metric.

The **Translation based metric** (Su and Babych, 2012a) uses MT APIs for document translation into English. Then four independent similarity feature functions are applied to a document pair:

- Lexical feature — both documents are pre-processed (tokenised, lemmatised, and stop-words are filtered) and then vectorised. The lexical overlap score is calculated as a cosine similarity function over the vectors of two documents.
- Structural feature — the difference of sentence counts and content word counts (equally interpolated).
- Keyword feature — the cosine similarity of top 20 keywords.
- NE feature — the cosine similarity of NEs (extracted using Stanford NER).

These similarity measures are linearly combined in a final comparability score. This is implemented by a simple weighted average strategy, in which each

type of feature is associated with a weight indicating its relative confidence or importance. The comparability scores are normalised on a scale of 0 to 1, where a higher comparability score indicates a higher comparability level.

The reliability of the proposed metric has been evaluated on a gold standard of comparable corpora for 11 language pairs (Skadiņa et al., 2010). The gold standard consists of news articles, legal documents, knowledge-base articles, user manuals, and medical documents. Document pairs in the gold standard were rated by human judges as being parallel, strongly comparable, or weakly comparable. The evaluation results suggest that the comparability scores reliably reflect comparability levels. In addition, there is a strong correlation between human defined comparability levels and the confidence scores derived from the comparability metric, as the Pearson R correlation scores vary between 0.966 and 0.999, depending on the language pair.

The **Dictionary based metric** (Su and Babych, 2012b) is a lightweight approach, which uses bilingual dictionaries to lexically map documents from one language to another. The dictionaries are automatically generated via word alignment using GIZA++ (Och and Ney, 2000) on parallel corpora. For each word in the source language, the top two translation candidates (based on the word alignment probability in GIZA++) are retrieved as possible translations into the target language. This metric provides a much faster lexical translation process, although word-for-word lexical mapping produces less reliable translations than MT based translations. Moreover, the lower quality of text translation in the dictionary based metric does not necessarily degrade its performance in predicting comparability levels of comparable document pairs. The evaluation on the gold standard shows a strong correlation (between 0.883 and 0.999) between human defined comparability levels and the confidence scores of the metric.

2.2 Parallel Sentence Extractor from Comparable Corpora

Phrase-based statistical translation models are among the most successful translation models that currently exist (Callison-Burch et al., 2010). Usually, phrases are extracted from parallel corpora by means of symmetrical word alignment and/or by phrase generation (Koehn et al., 2003). Our toolkit exploits comparable corpora in order to find and extract comparable sentences for SMT training using a tool named *LEXACC* (Ştefănescu et al., 2012).

LEXACC requires aligned document pairs (also *m* to *n* alignments) for sentence extraction. It also allows extraction from comparable corpora as a whole; however, precision may decrease due to larger search space.

LEXACC scores sentence pairs according to five lexical overlap and structural matching feature functions. These functions are combined using linear interpolation with weights trained for each language pair and direction using logistic regression. The feature functions are:

- a lexical (translation) overlap score for content words (nouns, verbs, adjectives, and adverbs) using GIZA++ (Gao and Vogel, 2008) format dictionaries;
- a lexical (translation) overlap score for functional words (all except content words) constrained by the content word alignment from the previous feature;
- the alignment obliqueness score, a measure that quantifies the degree to which the relative positions of source and target aligned words differ;
- a score indicating whether strong content word translations are found at the beginning and the end of each sentence in the given pair;
- a punctuation score which indicates whether the sentences have identical sentence ending punctuation.

For different language pairs, the relevance of the individual feature functions differ. For instance, the locality feature is more important for the English-Romanian pair than for the English-Greek pair. Therefore, the weights are trained on parallel corpora (in our case - 10,000 pairs).

LEXACC does not score every sentence pair in the Cartesian product between source and target document sentences. It reduces the search space using two filtering steps (Ştefănescu et al., 2012). The first step makes use of the Cross-Language Information Retrieval framework and uses a search engine to find sentences in the target corpus that are the most probable translations of a given sentence. In the second step (which is optional),

the resulting candidates are further filtered, and those that do not meet minimum requirements are eliminated.

To work for a certain language pair, *LEXACC* needs additional resources: (i) a GIZA++-like translation dictionary, (ii) lists of stop-words in both languages, and (iii) lists of word suffixes in both languages (used for stemming).

The performance of *LEXACC*, regarding precision and recall, can be controlled by a threshold applied to the overall interpolated parallelism score. The tool has been evaluated on news article comparable corpora. Table 1 shows results achieved by *LEXACC* with different parallelism thresholds on automatically crawled English-Latvian corpora, consisting of 41,914 unique English sentences and 10,058 unique Latvian sentences.

Threshold	Aligned pairs	Precision	Useful pairs
0.25	1036	39.19%	406
0.3	813	48.22%	392
0.4	553	63.47%	351
0.5	395	76.96%	304
0.6	272	84.19%	229
0.7	151	88.74%	134
0.8	27	88.89%	24
0.9	0	-	0

Table 1. English-Latvian parallel sentence extraction results on a comparable news corpus.

Threshold	Aligned pairs	Precision	Useful pairs
0.2	2324	10.32%	240
0.3	1105	28.50%	315
0.4	722	53.46%	386
0.5	532	89.28%	475
0.6	389	100%	389
0.7	532	100%	532
0.8	386	100%	386
0.9	20	100%	20

Table 2. English-Romanian parallel sentence extraction results on a comparable news corpus.

Table 2 shows results for English-Romanian on corpora consisting of 310,740 unique English and 81,433 unique Romanian sentences.

Useful pairs denote the total number of parallel and strongly comparable sentence pairs (at least 80% of the source sentence is a translation in the target sentence). The corpora size is given only as an indicative figure, as the amount of extracted parallel data greatly depends on the comparability of the corpora.

2.3 Named Entity Extraction and Mapping

The second workflow of the toolkit allows NE and terminology extraction and mapping. Starting with named entity recognition, the toolkit features the first NER systems for Latvian and Lithuanian (Pinnis, 2012). It also contains NER systems for English (through an *OpenNLP NER*[2] wrapper) and Romanian (*NERA*). In order to map named entities, documents have to be tagged with NER systems that support MUC-7 format NE SGML tags.

The toolkit contains the mapping tool *NERA2*. The mapper requires comparable corpora aligned in the document level as input. *NERA2* compares each NE from the source language to each NE from the target language using cognate based methods. It also uses a GIZA++ format statistical dictionary to map NEs containing common nouns that are frequent in location names. This approach allows frequent NE mapping if the cognate based method fails, therefore, allowing increasing the recall of the mapper. Precision and recall can be tuned with a confidence score threshold.

2.4 Terminology Mapping

During recent years, automatic bilingual term mapping in comparable corpora has received greater attention in light of the scarcity of parallel data for under-resourced languages. Several methods have been applied to this task, e.g., contextual analysis (Rapp, 1995; Fung and McKeown, 1997) and compositional analysis (Daille and Morin, 2008). Symbolic, statistical, and hybrid techniques have been implemented for bilingual lexicon extraction (Morin and Prochasson, 2011).

Our terminology mapper is designed to map terms extracted from comparable or parallel

[2] Open NLP - http://incubator.apache.org/opennlp/.

documents. The method is language independent and can be applied if a translation equivalents table exists for a language pair. As input, the application requires term-tagged bilingual corpora aligned in the document level.

The toolkit includes term-tagging tools for English, Latvian, Lithuanian, and Romanian, but can be easily extended for other languages if a POS-tagger, a phrase pattern list, a stop-word list, and an inverse document frequency list (calculated on balanced corpora) are available.

The aligner maps terms based on two criteria (Pinnis et al., 2012; Ştefănescu, 2012): (i) a GIZA++-like translation equivalents table and (ii) string similarity in terms of *Levenshtein* distance between term candidates. For evaluation, Eurovoc (Steinberger et al., 2002) was used. Tables 4 and 5 show the performance figures of the mapper for English-Romanian and English-Latvian.

Threshold	P	R	F-measure
0.3	0.562	0.194	0.288
0.4	0.759	0.295	0.425
0.5	**0.904**	**0.357**	**0.511**
0.6	0.964	0.298	0.456
0.7	0.986	0.216	0.359
0.8	0.996	0.151	0.263
0.9	0.995	0.084	0.154

Table 3. Term mapping performance for English-Romanian.

Threshold	P	R	F-measure
0.3	0.636	0.210	0.316
0.4	0.833	0.285	0.425
0.5	**0.947**	**0.306**	**0.463**
0.6	0.981	0.235	0.379
0.7	0.996	0.160	0.275
0.8	0.996	0.099	0.181
0.9	0.997	0.057	0.107

Table 4. Term mapping performance for English-Latvian.

3 Conclusions and Related Information

This demonstration paper describes the ACCURAT toolkit containing tools for multi-level alignment and information extraction from comparable corpora. These tools are integrated in predefined workflows that are ready for immediate use. The workflows provide functionality for the extraction of parallel sentences, bilingual NE dictionaries, and bilingual term dictionaries from comparable corpora.

The methods, including comparability metrics, parallel sentence extraction and named entity/term mapping, are language independent. However, they may require language dependent resources, for instance, POS-taggers, Giza++ translation dictionaries, NERs, term taggers, etc.[3]

The ACCURAT toolkit is released under the Apache 2.0 licence and is freely available for download after completing a registration form[4].

Acknowledgements

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347.

References

Sadaf Abdul-Rauf and Holger Schwenk. On the use of comparable corpora to improve SMT performance. EACL 2009: Proceedings of the 12th conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 16-23.

Sadaf Abdul-Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. Machine Translation, 25(4): 341-375.

ACCURAT D2.6 2011. Toolkit for multi-level alignment and information extraction from comparable corpora (http://www.accurat-project.eu).

Dan Gusfield. 1997. Algorithms on strings, trees and sequences. Cambridge University Press.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki and Omar Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, 17-53.

Béatrice Daille and Emmanuel Morin. 2008. Effective compositional model for lexical alignment. Proceedings of the 3rd International Joint Conference

[3] Full requirements are defined in the documentation of each tool (ACCURAT D2.6, 2011).
[4] http://www.accurat-project.eu/index.php?p=toolkit

on Natural Language Processing, Hyderabad, India, 95-102.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 440-447.

Pascale Fung and Kathleen Mckeown. 1997. Finding terminology translations from non-parallel corpora. Proceedings of the 5th Annual Workshop on Very Large Corpora, 192-202.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of a word alignment tool. Proceedings of ACL-08 HLT: Software Engineering, Testing, and Quality Assurance for Natural Language Processing, June 20, 2008. The Ohio State University, Columbus, Ohio, USA, 49-57.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), May 27-June 1, Edmonton, Canada.

Philip Koehn. 2010. Statistical machine translation, Cambridge University Press.

Bin Lu, Tao Jiang, Kapo Chow and Benjamin K. Tsou. 2010. Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT. Proceedings of the 3rd workshop on building and using comparable corpora: from parallel to non-parallel corpora, Valletta, Malta, 42-48.

Dragoș Ștefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from nonparallel corpora. ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Morristown, NJ, USA, 81-88.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. ACL HLT 2011, 27-34.

Mārcis Pinnis. 2012. Latvian and Lithuanian named entity recognition with TildeNER. Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, Tatiana Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June 20-21, Madrid, Spain.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. Proceedings of the 33rd annual meeting on Association for Computational Linguistics, 320-322.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. Proceedings of NAACL 2010, Los Angeles, USA.

Dan Ştefănescu. 2012. Mining for term translations in comparable corpora. Proceedings of the 5th Workshop on Building and Using Comparable Corpora (BUCC 2012) to be held at the 8th edition of Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, May 23-25, 2012.

Ralf Steinberger, Bruno Pouliquen and Johan Hagman. 2002. Cross-lingual document similarity calculation using the multilingual thesaurus Eurovoc. Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02), Springer-Verlag London, UK, ISBN:3-540-43219-1.

Inguna Skadiņa, Ahmet Aker, Voula Giouli, Dan Tufis, Rob Gaizauskas, Madara Mieriņa and Nikos Mastropavlos. 2010. Collection of comparable corpora for under-resourced languages. In *Proceedings of the Fourth International Conference Baltic HLT 2010*, IOS Press, Frontiers in Artificial Intelligence and Applications, Vol. 219, pp. 161-168.

Fangzhong Su and Bogdan Babych. 2012a. Development and application of a cross-language document comparability metric. Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

Fangzhong Su and Bogdan Babych. 2012b. Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-) parallel translation equivalents. Proceedings of EACL'12 joint workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), Avignon, France.

Dan Ştefănescu, Radu Ion and Sabine Hunsicker. 2012. Hybrid parallel sentence mining from comparable corpora. Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), Trento, Italy.