Computer Science Journal of Moldova, vol.**20**, no.2(59), 2012

Finding Translation Examples for Under-Resourced Language Pairs or for Narrow Domains; the Case for Machine Translation

Dan Tufiş

Abstract

The cyberspace is populated with valuable information sources, expressed in about 1500 different languages and dialects. Yet, for the vast majority of WEB surfers this wealth of information is practically inaccessible or meaningless. Recent advancements in cross-lingual information retrieval, multilingual summarization, cross-lingual question answering and machine translation promise to narrow the linguistic gaps and lower the communication barriers between humans and/or software agents. Most of these language technologies are based on statistical machine learning techniques which require large volumes of cross lingual data. The most adequate type of cross-lingual data is represented by parallel corpora, collection of reciprocal translations. However, it is not easy to find enough parallel data for any language pair might be of interest. When required parallel data refers to specialized (narrow) domains, the scarcity of data becomes even more acute. Intelligent information extraction techniques from comparable corpora provide one of the possible answers to this lack of translation data.

Keywords: alignment, comparable corpora, document crawling, machine learning, multilingual corpora, parallel corpora, statistical machine translation

1 Introduction

According to the Ethnologue site http://www.ethnologue.com/ there are 6909 languages in the world today. OLAC, the largest language

repository (The Open Language Archives Community) has linguistic data for 3930 languages. The Online Database of Interlinear text (ODIN) project (http://www.csufresno.edu/odin/) is a database of interlinear text "snippets" harvested mostly from scholarly documents posted on the Web (Lewis, 2006) and covers 1250 languages. Approximately 6% (389) of all extant languages are spoken by at least 1 million persons each, in total amounting for 94% of the Earth's population. Recent estimations (indigenoustweets.blogspot.com/2011/12/) approximate to 1500 the number of languages for which, on the web, one could find "primary texts": newspapers, blog posts, Wikipedia articles, Bible translations, etc. With such a linguistic diversity on the web, it is not surprising that the scientific and technological communities rank language technologies among the highest priorities. In a globalized information world, natural language communication mediated by computer is one of the most ambitious and difficult tasks. Cross-lingual information retrieval, natural question answering systems or machine translation are hot topics, substantially funded by international and national agencies. Large companies include these areas between their most promising research and development domains. Cross-lingual communication is not restricted to human use, but also it makes sense to conceive it among software agents, avatars of human users, in collaborative search for knowledge relevant to their masters' informational needs. Machine translation, probably the oldest scientific endeavor in computer science, is frequently called the Queen of Artificial Intelligence as it incorporates the majority of methods and techniques developed in various fields of AI and language engineering. In spite of more than 60 years of huge world-wide research and computational efforts to solve the problem, Machine Translation and Natural Language Understanding are still far away from the performances ascribed by Science Fiction Literature to humanoid robots. However, the last 10-15 years or so have seen huge scientific and technological progress in the area of natural language processing: "...specialized efficient language processing algorithms, hardware with greater computing power and storage capacities, large volumes of digitized text and speech data and, most importantly, powerful new methods of statistical language

processing that could exploit the language data for learning hidden regularities governing our language use. Lately Google's web search, Autonomy's text analytics, Nuance's speech technology, Google's on-line translation, IBM Watson's question answering and Apple Siri's personal assistance have given us but a glimpse of the massive potential behind the evolving language technologies. Leading-edge industry has already reacted, but this time much more decisively. IBM, SAP, SDL, Apple, Google, Amazon, Nokia, Nuance, Facebook and others have started acquiring language technology enterprises, among them many small promising start-up companies"[1]. Advanced technologies, enabled by natural language processors, such as those in automotive industry or, more recently in intelligent domotics, get us closer to the Science Fiction predictions, but only for very few languages. On-line machine translation services such as those offered by Google, Microsoft or Yahoo allow for assimilation of knowledge originally expressed in tens of languages unknown to the standard reader. Translations are available for many language pairs, but with very different quality. This is not arbitrary, but a direct consequence of the quantity and quality of the available linguistic resources.

2 Data hunger: better data is more data

With the great progress in data-driven machine learning methods and algorithms, to a large extent language independent, the focus of the natural language processing research shifted from individual language modelling to the more appealing multilingual statistical based approaches. The general lesson learnt from the latest research and development results is that the most urgent need is building an infrastructure to collect and distribute large quantities of multilingual data:

a) monolingual lexicons, grammars, text and speech corpora for as many languages as possible;

[1]Uszkoreit, H. (2012). Language Technology Before the Horizon. IT for Human Language, Understanding and Thought. Personal communication

b) bi-lingual lexicons, grammars, text and speech corpora for as many language pairs as possible.

Currently, there are several large international initiatives such as CLARIN-ERIC or META-NET which complement and improve the services of older language resources associations such as ELRA/ELDA in Europe or LDC in USA. They were funded for creating appropriate technological infrastructures to support:

a) collecting, organizing and disseminating information that gives an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programmes; prevailing standards and practices; current level of development, main drivers and roadblocks; etc;

b) assembling and preparing language resources for distribution. This includes collecting languages resources; documenting them and upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.

c) distributing the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaborating with other projects and, where useful, with other relevant multi-national forums or activities. This includes also help in building and operating broad inter-connected repositories and exchange facilities;

d) mobilising national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

The web is the largest space from where such multilingual data can be collected, but there are several impediments in using this data:

a) the web is highly unstructured and it requires significant effort to discover language data useful for technological developments;

b) the IPR restrictions on many valuable resources (mono- and bilingual dictionaries, lexical ontologies, literature, etc) prevent the downloading and further use for building language and translation models;

c) the IPR free resources are usually of low quality and need significant work to bring them to a usable quality; for instance, many texts in Romanian are written without diacritics, or with different character codes; user generated content is frequently affected by ungrammatical language, slang and coded abbreviations; additionally, the "google-isation" effect (posting on the web texts translated by Google) generates more and more poor quality language data. In spite of the data hunger of the statistical methods, the data source selection is of crucial importance in order to construct reliable language and translation models;

d) Quantities of useful language data (both monolingual and multilingual) is highly unballanced among the languages of the world.

The data-driven methods in machine translation among the language pairs for which large and good quality language resources exist (parallel corpora based on professional translations, bilingual electronic dictionaries, multilingual lexical ontologies, etc) demonstrated that the data issue is essential. Several experimental studies demonstrated that good quality automatic translations may be obtained in clearly delimited universes of discourse and for specific text registers, provided *enough* clean parallel text data is available. Typical examples are: formal language in juridical and legal area (e.g. the French-English parallel corpus based on The Hansards of the 36^{th} Parliament of Canada; the 22 language parallel corpus of the Acquis Communautaire) or instructional language as in user manuals (e.g. Microsoft, UNIX, KDE

manuals, etc). What "*enough*" parallel data means is dependent on the universe of discourse and the linguistic registers used in the corpus, but most successful translation systems report at least one million of sentence pairs in the training/learning translation models and over one billion words in the monolingual corpora used for language modelling. To have a rough idea, a novel such as famous Orwell's "1984" has about 6400 sentences per language (and about 110,000 words) in the Multext-East parallel corpus containing a dozen of different translations of the original.

On the Web there are almost 400 languages, each being spoken by more than 1,000,000 persons. It might sound as a reasonable objective to develop cross-lingual technological services for any pair of these languages. That is, almost 80,000 uni-directional language pairs! Theoretically, such an aim may be attained with the cutting edge technologies based on statistical methods and machine learning. The problem is that it is impossible to find on the Web parallel corpora for more than 200-300 of these language pairs. In fact, except maybe for less than a dozen of languages, the extant parallel corpora are very small, highly specialized or unavailable for research and development purposes.

Is this a dead-end? No, because following the lead of Munteanu and Marcu (2005), recent research (Rauf and Schwenk, 2009, 2011), (Ion et al, 2011a), (Skadiņa et al, 2010a,b; 2012), (Ştefănescu et al., 2012) etc., developed very promising methods to overcome this gap, by mining large bi-lingual collections of *comparable documents.* Unlike parallel data, comparable data can be found on the Web in much larger quantities (several orders of magnitude). Such collections are referred to *as comparable corpora.* A pair of documents is called comparable if they are about the same topic and use fragments of text that might be considered reciprocal translations. Based on this definition, depending on the quantity of overlapping translations, comparable corpora may be classified as *strongly comparable, medium comparable* and *weakly comparable.*

Comparable corpora for a language pair L1-L2 are usually built based on focused collection of monolingual data in L1 and L2 followed by a preliminary pairing of the cross-lingual most similar documents

in the two collections of documents. Afterwards, the document pairs with the similarity scores above a user-selected threshold are subject to an in-depth analysis to detect any parallel or almost parallel sentences. The major processing steps are suggested by the diagram in Figure 1 and will be briefly described in the following sections.

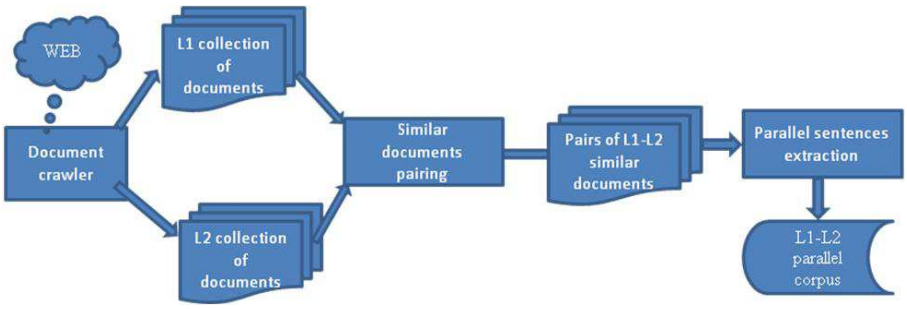

Figure 1. Processing flow for extraction of parallel data from comparable corpora

3 Collecting general and domain specific comparable corpora

The recently finished ACCURAT project[2] (2010-2012) developed several innovative methods and efficient algorithms to collect comparable data (Skadiņa et al, 2012). In principle, there are two types of comparable corpora one would be interested in collecting from the Web: general language corpora and domain specific corpora, containing more often than not specialized terminology.

Aker et al. (2012) describe the methodology and assumptions used to collect general language texts and assemble them into corpora. To do this, they make use of the current news articles and download huge amount of article titles using Google News Search and RSS News feeds. The downloaded titles are split into different bins based on their pub-

[2]European project no. 248347 (www.accurat-project.eu).

lication dates. Each bin contains titles of the same week. Based on the motivation that news titles are a good indicator for the content of the news document (Edmundson, 1969, Lopez et al. 2011) the titles from each bin are taken as representatives of document contents and paired using different heuristics such as cosine similarity, title length difference, and publication date difference. Only contents of "good" article pairs are downloaded. This reduces costs measured in hard disk space and computational power, and also reduces noise in the pairing process by limiting search to one week time span. Following this strategy, for the languages of the project (Croatian, English, Estonian, German, Greek, Latvian, Lithuanian, Romanian and Slovene) the general crawler collected and preliminary classified as mentioned above tens of thousands of documents. The number of harvested and paired documents varied from 720 for Croatian-English to 29341 for German-English, with an average of 7366 documents per language pair. The document pairing based on the general heuristics (e.g. title cosine similarity, publication data) is extremely fast but inherently imprecise. However, the reduction of search space is significant, making room for more sophisticated and time-consuming algorithms to further filter out as much noise as possible.

For collecting domain-specific corpora from the web, a highly configurable Focused Monolingual Crawler (FMC) tool has been developed by our Greek partners from ILSP, based on the Bixo[3] open-source web mining toolkit. Given a narrow domain (topic) and a language, the FMC tool requires two manually or semi-automatically produced input datasets: (i) a list of topic multi-word term expressions and (ii) a list of topic-related seed URLs (Skadiņa et al, 2012).

The user can then optionally configure FMC in a variety of ways, e.g. set file types to download, domain filtering options, self-terminating conditions, crawling politeness parameters, to name but a few. Crawling starts from the seed URLs and expands dynamically to other URLs, while a lightweight text classification is performed on the web pages being visited, so that to retrieve only those web documents that are relevant to the chosen topic. Operations such as boilerplate

[3]http://bixo.101tec.com/

removal, text normalization and cleaning, language identification, etc. are done during runtime, whereas some post-crawling processing steps (including removing duplicates, post-classification and filtering, etc.) are also implemented. The FMC output consists of the collected web documents in HTML and text format (UTF-8 encoding) as well as their metadata. Similarly to general language crawler, FMC achieves a preliminary pairing, removing from further consideration as many as possible unrelated documents. By using FMC, 28 comparable corpora have been constructed on 8 narrow domains[4], in 6 language pairs[5] amounting to a total of more than 148M tokens.

4 Pairing similar documents in comparable corpora

A metric for measuring comparability of pairs of documents in different languages performs two main functions: (1) *evaluates the quality* of the collected comparable corpora (2) *enhances* the corpora by ranking pairs of documents by their comparability, which indicates the likelihood of retrieving good-quality translation equivalents from the aligned document pairs. The consortium developed several programs to evaluate the comparability of the documents in the collected corpora.

Each of these programs generates pairs of documents associated with a comparability score. They differ both in accuracy but also in the running time necessary to complete the task. For instance, the EMACC (Expectation Maximization Alignment for Comparable Corpora) tool (Ion et al, 2011) although provided almost perfect pairings the algorithm is highly intensive and needs several days to finish pairing a relatively small comparable corpus of around 5000 documents per language.

On the other hand, the lexical based metric, DicMetric (Su et al., 2011) does not make categorical decisions and computes (very fast)

[4]Renewable Energy, Political News, Sports News, Technological News, Natural Disasters, Automotive Engineering, Assistive Technology, Software Localization

[5]EN-LV, EN-LT, EN-HR, EN-RO, EN-EL and EN-DE

similarity scores among various pairs of documents. It is based on bilingual dictionaries, and uses lexical, keyword, and named entity features which are weighted and compared as cosine similarity between the feature vectors. The weights were experimentally established so that the combination of these internal features could accurately predict externally defined comparability categories: parallel corpora, strongly comparable corpora and weakly comparable corpora. DicMetric produces a number in the range [0, 1], with higher values corresponding to greater comparability. After the pairing analysis, the [0, 1] interval was split into three intervals [0.1, 0.2), [0.2, 0.4), and (0.4, 1] corresponding to the weakly comparable, strongly comparable and parallel documents as judged by the human assessors. The Spearman correlation among the automatic labeling and human annotation was very high, ranging from 0.883 to 0.999 with an average of 0.975.

5 Extraction of MT-related data from comparable corpora

By "MT-related data" extracted from comparable corpora we understand collections of translation equivalent chunks of text. Such a chunk may contain a pair of terminological expressions, a pair of named entities, a pair of regular phrases or even a pair of sentences or paragraphs. The ACCURAT project developed several tools for extracting this kind of translation equivalents. The general approach for less-resourced pairs of languages was to first extract monolingually lists of name entities and terms for each project language, and then to map crosslingually the extracted lists. All these tools are largely documented in one public deliverable of the project (Ion et al., 2011b) and can be downloaded from the project's public site http://www.accurat-project.eu. For name entities extraction the consortium partners either reused and adapted public language independent software which comes al-

ready trained for English (OpenNLP[6], Stanford NER[7] and MENER[8]), trained for some project languages (Latvian Lithuanian, Greek) or developed new tools, language dependent due to specific tokenization rules (NERA1 for English and Romanian, CroNerc for Croatian). The underlying processing models vary from Conditional Random Fields (Stanford NER), Maximum Entropy (OpenNLP, MENER, NERA1) to Rule-based (CroNerc). The monolingual terminological and named entities lists were cross-lingually mapped using GIZA++ dictionaries and various string-similarity measures (Stefănescu, 2012).

Identifying corresponding named entities in different languages works reasonably well (with precision better than 90%). However, this is not the case for mapping technical terms. We argue that one of the reasons is the lack of terminological and name-entity gold-standards and as such, the interpretations are highly subjective.

The extraction of chunks of parallel phrases and sentences from comparable corpora is a more difficult task than extraction and cross-lingual mapping of named entities and terms. The usual sentence alignment techniques applicable for parallel corpora rely on a fundamental property: the translation equivalent paragraphs (and to a large extent, sentences) have the same order in the two parts of the bitext. This property, which significantly reduces the alignment search space, is not valid anymore in comparable corpora.

LEXACC is a Lucene[9]-based phrase extraction algorithm from comparable corpora (Ştefănescu et al., 2012) using cross-lingual information retrieval techniques. This program has been designed and implemented with the main emphasis on weakly comparable documents and, when available, it uses document pairing which could be explicitly specified (ex EMACC provided) or take into account all document pairs with a comparability score above a user specified threshold (as DicMetric generates). If document pairing is not available, it overcomes this lack

[6]http://incubator.apache.org/opennlp/index.html,

[7]http://nlp.stanford.edu/software/CRF-NER.shtml

[8]A highly modified version of the system developed by Chieu and Ng, the best-scoring system in the CoNLL-2003 shared task.

[9]http://lucene.apache.org/core/

on the expense of significant additional running time.

The collected documents in one of the languages of the comparable corpus (let's call it the target language) are multi-criterially indexed using the Lucene environment. The indexing phase requires a light pre-processing step[10]: each sentence of the target documents is stemmed (a list of endings in each language of interest is necessary) and all the functional words are removed (a list of functional words in each language of interest is required). Besides the stemmed content words, each target sentence is indexed by *its length class* (short, long and average) and the document pair in which it appears. The *length class* of an indexed sentence is computed based on average lengths of indexed sentences and standard deviation of the current sentence length from the average length. After the indexing phase, each sentence in the source language documents is turned into a Lucene Boolean query. This query generation follows the following steps (Ştefănescu et al., 2012):

1) the input sentence is stemmed and the functional words are purged;

2) the remaining stems are replaced by a disjunction of translation equivalents (a translation dictionary for the language pair of interest is necessary);

3) the query is conjunctively added the estimated length class of the parallel sentence (short, long and average);

4) if available, the document pair which the source sentence belongs to is added as a search constraint.

[10]Actually, LEXACC may take advantage of "heavier" resources (large bilingual lexicons) and processing tools: part of speech tagging, lemmatization, chunking. Indexing and retrieval of the information of interest is much more precise if such pre-processing is available. However, in the standard version we opted for relying on minimal pre-processing tools (stemmer) and resources (lists of typical endings for inflective languages, lists of functional words and seed bilingual lexicons) so that any pair of languages (especially for under-resourced ones) could be dealt with.

The translation dictionaries used in step 2) are automatically extracted using GIZA++ (Och and Ney, 2003) from whatever parallel corpora may be found for the considered language pair. The better translation dictionaries the better extraction results are.

The query built this way, is sent to Lucene search engine and the best matching N target sentences (implicitly 50) are returned. If the pairing information is available (item 4 above) the search will be restricted only in the target document of the pair. Otherwise, the search will consider all indexed sentences in the target language.

All the target sentences returned by Lucene search engine are similarity scored against the source sentence. LEXACC reuses the similarity measure of RACAI's previous PEXACC system (Ion, 2011) which is a weighted sum of several *reifying* feature values (Tufiş et al., 2006). These features have been selected (Ştefănescu, et al. 2012), as indicative for the "parallelism" of two sentences: translation probabilities, relative position of the translation equivalents, final punctuation, etc. The weights are dependent on language pair and were optimized by a logistic regression classifier (trained on 10,000 parallel sentences as positive examples and 10,000 non-parallel sentences as negative examples. For a detailed presentation of LEXACC see (Ştefănescu et al., 2012). Besides full sentences, LEXACC may extract sub-sentential fragments as well. In this case the size of extracted data is significantly larger. Because manual validation is a very time consuming task, we restricted ourselves only to parallel sentence pair evaluation. Another motivation for preferring parallel sentences to sub-sentential chunks stems from the need to avoid as much as possible duplication. Identical sentence pair, although present in comparable corpora, are less numerous than sub-sentential word groups. All the sentence pairs that receive a similarity score above a user established confidence threshold, are retained and added to the parallel corpus under construction.

The quantity of (quasi-)parallel sentences extracted from the news corpora collected by the consortium for each language pair varies depending on the quantity of crawled documents and their comparability scores.

The Table 1 shows, for four language pairs, the results of manual

evaluation of the extraction process. As one can see, the extraction rate for English-Latvian comparable collected corpus is 1.44% with a precision of 84%. It means that 84% of extracted sentence pairs were correct and that for extracting 1 Mb of parallel text, 69.4 Mb of comparable corpus had to be collected. The table also shows, that when confidence threshold is increased (from 0.27 to 0.45) the precision of the extracted data also increased. As a matter of fact, we found that for a confidence threshold of 0.6 all the English-Romanian extracted sentence pairs were correct.

Table 1. Parallel sentences extracted by LEXACC from the ACCURAT News Comparable corpora

Lang. pair	Size of comparable corpora	# Extracted sentence pairs/size (MB)	Confidence threshold	Precision
en-lv	76.34 MB	3679 /1.1MB	0.27	84%
en-lt	74.80 MB	1583 /0.57MB	0.27	84%
en-et	34.78 MB	673 /0.2 MB	0.27	84%
en-ro	71 MB	2019 / 0.6MB	0.45	93%

This evaluation allows us to answer a possible question (at least with respect to the analyzed languages): what is the quantity of comparable corpora one has to collect in order to get *enough* data for a machine translation experiment? The answer depends on the language pair, the domain of the comparable corpora and what the purpose of the data is. If the data is meant for building a genuine new SMT, the answer is related to what we said in the section 2. If we want an English-Latvian SMT for the news domain, we would need to collect about 20.7 GB of comparable texts out of which we will presumably extract more than 1,000,000 of parallel sentences. If the extracted parallel data will be used in domain adaptation of an existing English-Latvian SMT, the size of the necessary comparable news corpora would

be around 0.8 GB[11]. According to the current technology these figures are not scarring anymore. If such data is somewhere on the Web it may be put to good service for lesser resource languages. These estimations took into account some kind of worst-case scenario, because from the point of view of potential parallel sentences, the News corpora collected by the consortium with a general crawler may be characterized as weakly comparable corpora. For strongly comparable corpora, such as Wikipedia, the parallel sentence extraction rate is much higher (e.g. for Romanian-English this rate was about 70%, that is 80 times higher than for the News comparable corpora). In general, state of the art focused crawlers produce comparable corpora with much higher degree of comparability than the general crawlers, but on the other hand, they collect less data. The Table 2 and Table 3 exemplify the monolingual corpora collected in a narrow domain – renewable energy (Table 2) – from which LEXACC extracted the (quasi-)parallel sentences (Hunsicker and Chen, 2012) for various language pairs.

Table 2. Collected comparable monolingual corpora about renewable energy

LANGUAGE	**SIZE (SENTENCES)**
Croatian	19,742
Lithuanian	62,902
Latvian	23,893
Romanian	39,671
English	607,816

As one can see, the vast majority of the Latvian documents were translation of some English documents and to a large extent this was also the case for Romanian documents.

The extracted parallel data was used to adapt to the new domain (renewable energy) some general baseline SMT systems for the lan-

[11]This estimation is based on experiments conducted by Sabine Hunsicker of DFKI within the ACCURAT project, which showed that good results in domain adaptation of a reliable SMT would require about 40,000 domain specific parallel sentences.

Table 3. Parallel data extracted from renewable energy comparable corpora

LANGUAGE PAIR	**SIZE (SENTENCES)**	**EXTRACTION RATE (%)**
Croatian-English	8,237	41.72
Lithuanian-English	16,743	26.61
Latvian-English	22,992	96.22
Romanian-English	26,939	67.90

guage pairs shown in Table 3. The improvements of the translation quality, measured in terms of BLEU scores (Hunsicker and Chen, 2012) were significant, ranging from 3.04 point for English-Romanian up to 31.84 points for English Lithuanian.

6 Conclusions

In this article we described a processing flow for exploiting comparable corpora in collecting parallel sentences meant for improving translation quality for under resourced languages and/or narrow domains. We presented tools and resources for collecting, evaluating and aligning of comparable texts for application in machine translation. MT-related data extracted from comparable corpora (parallel named entities pairs, parallel term pairs, parallel sub-sentential chunks and parallel sentences) can be reliably found even in weakly comparable corpora. Given that comparable corpora can be collected in large quantities (say GB), even a few percentages of extracted MT-related data can provide a significant help in building or adapting a SMT for which proper training parallel corpora cannot be easily found.

Tools and resources described in this paper are publicly available and largely described on ACCURAT project website: www.accurat-project.eu.

Acknowledgments. This work has been supported by the AC-

CURAT project (www.accurat-project.eu/) funded by the European Community's Seventh Framework Program (FP7/2007-2013) under the Grant Agreement n° 248347, and co-financed by the national research program Capacities under the grant 72EU/11.06.2010

References

[1] Aker, A., Kanoulas, E. and Gaizauskas, R. (2012). *A light way to collect comparable corpora from the Web.* In Proceedings of LREC 2012, 21-27 May, Istanbul, Turkey.

[2] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm.* In Journal of the Royal Statistical Society, 39(B): pp. 1–38.

[3] Hunsicker S., Chen Y. (2012). *ACCURAT Deliverable 4.3: Improved baseline SMT systems adjusted for narrow domain.* 1^{st} of March 2012 (www.accurat-project.eu).

[4] Ion, R., Ceauşu A. and Irimia, E. (2011a). *An Expectation Maximization Algorithm for Textual Unit Alignment.* In Proceedings of 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, pp. 128–135.

[5] Ion, R. Pinnis, M., Ştefănescu, D., Aker, A., Paramita, M., Su, F., Irimia, E., Zhang, X., Radu Ion, Mrcis Pinnis, Dan Ştefănescu, Ahmet Aker, Monica Paramita, Fangzhong Su, Elena Irimia, Xiaojun Zhang, Ljubešić, N. (2011b). *Toolkit for multi-level alignment and information extraction from comparable corpora.*, 31^{st} August (http://www.accurat-project.eu/), 123 pages.

[6] Ion, R. (2012) *PEXACC: A Parallel Data Mining Algorithm from Comparable Corpora.* In Proceedings of LREC 2012, 21-27 May, Istanbul, Turkey.

[7] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C.,

Bojar, O., Constantin, A. and Herbst, E. (2007). *Moses: open source toolkit for statistical machine translation.* In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 177–180.

[8] Lewis W., D., (2006) *ODIN: A Model for Adapting and Enriching Legacy Infrastructure.* Proceedings of the Workshop on e-Humanities – An Emerging Area of Concern, Second International Conference on e-Science and Grid Technologies (e-Science 2006), 4-6 December 2006, Amsterdam, The Netherlands. IEEE Computer Society 2006, ISBN 0-7695-2734-5 .

[9] Munteanu, D. and Marcu, D. (2005). *Improving Machine Translation Performance by Exploiting Non-Parallel Corpora.* Computational Linguistics, 31(4), pp.477–504.

[10] Och, F. J., Ney, H. (2003). *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, volume 29, number 1, pp. 19–51.

[11] Rauf, S.A., Schwenk, H. (2009) *On the use of comparable corpora to improve SMT performance.* In: EACL 2009: Proceedings of the 12th conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, pp 16–23.

[12] Rauf, S.A., Schwenk, H. (2011). *Parallel sentence generation from comparable corpora for improved SMT.* In: Machine Translation, 25(4), pp.341–375.

[13] Skadiņa, I., Aker, A., Giouli, V., Tufis, D., Gaizauskas, R., Mieriņa M. and Mastropavlos, N. A. (2010a). *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.

[14] Skadiņa, I., Vasijevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D, Gornostay, T. (2010b). *Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation.* In

Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, European Language Resources Association (ELRA), La Valletta, Malta, May 2010, pp. 6–14.

[15] Skadiņa, I., Aker, A., Mastropavlos,N., Su, F., Tufiş, D., Verlic, M., Vasiljevs, A., Babych, B., Glaros,N. (2012). *Collecting and Using Comparable Corpora for Statistical Machine Translation.* In Proceedings of LREC 2012, 21-27 May, Istanbul, Turkey.

[16] Su, F., Babych, B., Paramita, M., Gaizauskas, R. (2011). *ACCURAT Deliverable 1.3: Evaluation and Elaboration of metrics,* December, (www.accurat-project.eu).

[17] Ştefănescu, D. (2012). *Mining for Term Translations in Comparable Corpora.* Proceedings of Building and Using Comparable Corpora, May 26^{th},Istanbul, Turkey.

[18] Ştefănescu, D., Ion, R., and Hunsicker, S. (2012). *Hybrid Parallel Sentence Mining from Comparable Corpora.* Proceedings of European Association for Machine Translation, Trento, Italy, 28-30 May, pp 137–144.

[19] Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2006). *Improved Lexical Alignment by Combining Multiple Reified Alignments.* In Proceedings of the 11^{th} Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3-7 April, pp. 153-160, ISBN 1-9324-32-61-2.

[20] Tufiş, D., Dumitrescu, Ş. (2012). *Cascaded Phrase-Based Statistical Machine Translation Systems.* Proceedings of European Association for Machine Translation, Trento, Italy, 28-30 May, pp. 129–136.

Dan Tufiş,

Received June 28, 2012

Research Institute for Artificial Intelligence
Romanian Academy
13, "13 Septembrie", 050711, Bucharest 5, Romania
E–mail: *tufis@racai.ro*