Inguna Skadiņa, Andrejs Vasiļjevs, Raivis Skadiņš, Robert Gaizauskas, Dan Tufiş, and Tatiana Gornostay. *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" (BUCC10) at the 7th Language Resources and Evaluation Conference (LREC 2010)*, pp. 6-14, Valletta, Malta, May 2010.

Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation

Inguna Skadiņa*, Andrejs Vasiļjevs*, Raivis Skadiņš*, Robert Gaizauskas†, Dan Tufiş‡ and Tatiana Gornostay*

* Tilde, Riga, Latvia
† Department of Computer Science, University of Sheffield, Sheffield, UK
‡ Research Institute for Artificial Intelligence, Romanian Academy Bucharest, Romania
inguna.skadina@tilde.lv, andrejs@tilde.lv, raivis.skadins@tilde.lv, r.gaizauskas@sheffield.ac.uk,
dan_tufis2006@yahoo.com, tatiana.gornostay@tilde.lv

Abstract

Lack of sufficient linguistic resources and parallel corpora for many languages and domains currently is one of the major obstacles to further advancement of automated translation. The solution proposed in this paper is to exploit the fact that non-parallel bi- or multilingual text resources are much more widely available than parallel translation data. This position paper presents previous research in this field and research plans of the ACCURAT project. Its goal is to find, analyze and evaluate novel methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for under-resourced languages and narrow domains.

1. Introduction

In recent decades data-driven approaches have significantly advanced the development of machine translation (MT). However, the applicability of current data-driven methods directly depends on the availability of very large quantities of parallel corpus data. For this reason the translation quality of current data-driven MT systems varies dramatically from being quite good for language pairs with large corpora available (e.g. English and French) to being barely usable for under-resourced languages and domains (e.g. Latvian and Croatian).

The problem of availability of linguistic resources is especially relevant for "smaller" or under-resourced languages. For example, one of the few parallel corpora of reasonable size for Latvian is the JRC Acquis corpus (Steinberger et al, 2006) which contains EU legislation texts. SMT trained on this corpora performs well on EU legislation documents (Koehn et al, 2009; Skadiņa and Brālītis, 2009), but it has unacceptable results for other domains.

The solution proposed in ACCURAT project and presented in this paper is to exploit the fact that comparable corpora, i.e., non-parallel bi- or multilingual text resources are much more widely available than parallel translation data.

Comparable corpora have several obvious advantages over parallel corpora – they can draw on much richer, more available and more diverse sources which are produced every day (e.g. multilingual news feeds) and are available on the Web in large quantities for many languages and domains. Although the majority of these texts are not direct translations, they share a lot of common paragraphs, sentences, phrases, terms and named entities in different languages. Expansion of Web content and massive library digitization initiatives make comparable corpora much more available than parallel corpora

The FP7 ACCURAT (Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation) project has started on January 1, 2010. The main goal of this 2.5 year project is to find, analyze and evaluate novel methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for under-resourced languages and narrow domains.

A typical statistical MT system is based on a language model trained on monolingual target language corpus, and a translation model. Methods for the creation of translation models from parallel corpora are well studied and there are several techniques developed and widely available.

However, similar methods and techniques for non-parallel, or comparable corpora, have not been worked out thoroughly and there has been relatively little research on this subject.

This position paper presents research plans of the ACCURAT project to create a methodology and fully functional model for exploiting comparable corpora in MT, including corpus acquisition from the Web and other sources, analysis and metrics of comparability, multi-level alignment and extraction of lexical data and techniques for applying aligned text and extracted lexical data to increase the translation quality of existing MT systems.

The paper describes the state of the art in research related to use of comparable corpora for MT, presents related work regarding MT strategies and corpora use in MT, and describes the ACCURAT project goals and planned innovation.

2. Related Work in Corpus Use in MT

2.1 MT Strategies

Several approaches are used in the development of translation technologies: rule-based, statistical and example-based approaches. Cost-effectiveness is one of

the key reasons that the statistical paradigm has come to be the dominant current framework for MT theory and practice, as it has proven to be the most effective solution both from the point of view of time and labor resources and translation output quality.

Statistical Machine Translation (SMT) started with word-based models, but significant advances have been achieved with the introduction of phrase-based models (Koehn et al., 2003). Currently the most competitive SMT systems use phrase translation, such as the ATS (Och and Ney, 2004), CMU (Vogel et al., 2003) and IBM (Tillmann, 2003) systems. Recent work has also incorporated syntax or quasi-syntactic structures (Chiang, 2007). There are efforts to integrate in SMT systems linguistic annotation either at the word-level with factored translation models (Koehn and Hoang, 2007) or using tree-based models (Yamada and Knight, 2001, 2002; May and Knight, 2007). The proposed methods improve MT performance especially for languages with rich morphology and free word order, and help to solve such problems as long distance reordering and sentence-level grammatical coherence.

Until recently SMT research has been mainly focused on widely used languages, such as English, German, French, Arabic, and Chinese. For "smaller" languages MT solutions, as well as language technologies in general, are not as well developed due to the lack of linguistic resources and technological approaches that enable MT solutions for new language pairs to be developed cost effectively. This has resulted in a technological gap between these two groups of languages.

Although in the past few years translation services like Google Translate have started to broaden the set of translation language pairs, incorporating, e.g. the Baltic languages, translation quality lags behind significantly compared to major language pairs.

Also the EuroMatrix project[1] represents a major push in MT technology, applying the latest MT technologies systematically to all pairs of EU languages. The EuroMatrixPlus project[2] is continuing the rapid advance of MT technology, creating sample systems for every official EU language. Still these services and projects rely on available parallel corpus data.

2.2. Corpora Use in MT

In the area of rule-based MT systems, approaches towards using corpus-based technology for bilingual term extraction, and importing such terms into the dictionary of a rule-based system have been researched (Eisele et al., 2008).

Changes in the MT engine's process of data-driven term selection in the transfer component show that disambiguation of transfer alternatives can be significantly improved using the corpus-based data-driven techniques (Thurmair, 2006).

While SMT techniques are language independent, they require very large parallel corpora for training translation models. Translation systems trained on data from a particular domain, e.g. parliamentary proceedings, will perform poorly when used to translate texts from a different domain, e.g. news articles (Munteanu et al., 2004).

Parallel corpora remain a scarce resource covering few language pairs with too little data in only a few domains. For smaller languages parallel corpora are very limited in quantity, genre and language coverage. This remains true despite the creation of automated methods to collect parallel texts from the Web (Goutte et al., 2009; Hewavitharana and Vogel, 2008; Maia and Matos, 2008; Alegria et al., 2008; Munteanu, 2006; Munteanu and Marcu, 2005; Resnik and Smith, 2003).

The ACCURAT project goal is to overcome the bottleneck of insufficient parallel corpora for less widely used languages by extracting linguistic data from comparable corpora. Such corpora can be obtained by taking advantage of existing methods for mining the Web for similar documents or by other methods that will be explored in the project, such as mining Wikipedia.

3. Comparable Corpora

A comparable corpus is a relatively recent concept in MT, corpus linguistics and NLP in general. In contrast to the notion of a parallel corpus, a comparable corpus can be defined as collection of similar documents that are collected according to a set of criteria, e.g. the same proportions of texts of the same genre in the same domain from the same period (McEnery and Xiao, 2007) in more than one language or variety of languages (EAGLES, 1996) that contain overlapping information (Munteanu and Marcu, 2005; Hewavitharana and Vogel, 2008).

Examples of comparable corpora are:

- Comparable multilingual Document Collection in the Multilingual Corpora for Cooperation includes financial newspaper articles from the early 1990s in six European languages: Dutch (8.5 million words), English (30 million words), French (10 million words), German (33 million words), Italian (1.88 million words), and Spanish (10 million words).
- Bulgarian-Croatian comparable corpus (Bekavac et al. 2004) in news domain: 3,500,000 tokens (393 Kw Bulgarian; 3.1 Mw Croatian) was built from subsets of two larger newspaper corpora of respective languages from the texts selected using the same criteria (e.g., identical year, same domain etc.);
- English-Finnish-Swedish comparable corpus in news domain (University of Tampere);
- English-French-Norwegian comparable corpus in science domain (academic prose), 450 reviewed scientific papers; 3,2 million words;
- project INTERA and its four parallel sub-corpora.

These comparable corpora cannot be readily used for MT and are restricted to particular languages and certain domains. The degree of comparability of these corpora varies significantly, since texts were selected on the basis of one criterion only – topic.

[1] http://www.euromatrix.net

[2] http://www.euromatrixplus.net

Research in comparable corpora started about 15 years ago with the first works on general lexica (Rapp, 1995) and named entity translation derived from *noisy parallel corpora* (Fung, 1995). The authors supposed that the quantity of training data has an impact on the performance of statistical machine translation and a comparable corpus can compensate for the shortage of parallel corpora. This has been confirmed by other recent experiments (Munteanu and Marcu, 2005; Maia and Matos, 2008; Hewavitharana and Vogel, 2008; Goutte et al., 2009)

The latest research has also shown that adding extracted aligned parallel lexical data (additional phrase tables and their combination) from comparable corpora to the training data of an SMT system improves the system's performance in view of un-translated word coverage (Hewavitharana and Vogel, 2008). It has been also demonstrated that language pairs with little parallel data are likely to benefit the most from exploitation of comparable corpora. Munteanu (2006) achieved performance improvements of more than 50% from comparable corpora of BBC news feeds for English, Arabic and Chinese over a baseline MT system, trained only on existing available parallel data. The authors stated that the impact of comparable corpora on SMT performance is "comparable to that of human-translated data of similar size and domain".

One of the most challenging tasks is to perform alignment of comparable corpora for extraction of necessary translation data. Zhao and Vogel (2002) and Utiyama et al. (2003) extended algorithms designed to perform sentence alignment of parallel texts to apply them for comparable corpora. They started by attempting to identify similar article pairs from the two corpora. Then they treated each of those pairs as parallel texts and aligned their sentences by defining a sentence pair similarity score and use dynamic programming to find the least-cost alignment over the whole document pair. The performance of these approaches depends heavily on the ability to reliably find similar document pairs. Moreover, comparable article pairs, even those similar in content, may exhibit great differences at the sentence level (reordering, additions, etc). Therefore, they pose hard problems for the dynamic programming alignment approach.

The STRAND Web-mining system by Resnik and Smith (2003) can identify translational pairs. However, STRAND focuses on extracting pairs of parallel Web pages rather than sentences.

Muntenau and Marcu (2005) proposed a maximum entropy classifier that, given a pair of sentences, can determine whether or not they are translations of each other. This approach supposedly overcomes some of the limitations of previous approaches. Their experiments were carried out on Chinese, Arabic, and English non-parallel newspaper corpora.

ACCURAT will investigate previous multi-level alignment methods and will work on a complex approach to extract maximum linguistic data from comparable corpora for a number of under resourced languages (Croatian, Estonian, Greek, Latvian, Lithuanian and Romanian) and narrow domains. In such a way we will continue from a point reached by previous research.

4. ACCURAT Project

The main goal of the ACCURAT research is to find, analyze and evaluate novel methods how comparable corpora can compensate for this shortage of linguistic resources to improve MT quality significantly for under-resourced languages and narrow domains. Thus the project has the following key objectives:

- To create comparability metrics, i.e., to develop the methodology and determine criteria to measure the comparability of source and target language documents in comparable corpora.
- To develop, analyze and evaluate methods for automatic acquisition of comparable corpora from the Web.
- To elaborate advanced techniques for extraction of lexical, terminological and other linguistic data (e.g., named entities) from comparable corpora to provide training and customization data for MT.
- To measure improvements from applying acquired data against baseline results from SMT and RBMT systems.
- To evaluate and validate the ACCURAT project results in practical applications.

We will use the latest state-of-the-art in SMT and rule-based MT systems as a baseline and will provide novel methods to achieve much better results by extending these systems through the use of comparable corpora. Initial research demonstrates promising results from the use of comparable corpora in SMT (Munteanu and Marcu, 2005) and RBMT (Thurmair, 2006) and this makes us confident of the feasibility of the proposed approach.

The ACCURAT target is to achieve strong improvement in translation quality for a number of new EU official languages and languages of associated countries (Croatian, Estonian, Greek, Latvian, Lithuanian and Romanian), and propose novel approaches for adapting existing MT technologies to specific narrow domains, significantly increasing language and domain coverage of automated translation.

4.1. Comparability Metrics

The issue of comparability of corpora can be traced back to the origin of large-scale corpus research, when the aim was to balance the composition of a corpus to achieve representativeness (Sinclair, 1987). However we still lack definite methods to determine the criteria of comparability and comparability metrics to evaluate corpus usability for different tasks, such as machine translation, information extraction, cross-language information retrieval.

4.1.1. Criteria of Comparability and Parallelism

Comparability and parallelism is a complex issue, which can be applied to different levels, such as

- document collections,

- individual documents,
- paragraphs or sentences of documents.

Until now there has been no agreement on the degree of similarity that documents in comparable corpora should have, or even agreement about the criteria for measuring parallelism and comparability. There are only a few publications discussing the characteristics of comparable corpora (Maia, 2003). There have been some attempts to determine different kinds of document parallelism in comparable corpora, such as complete parallelism, noisy parallelism and complete non-parallelism, and define criteria of parallelism of similar documents in comparable corpora, such as similar number of sentences, sharing sufficiently many links (up to 30%), and monotony of links (up to 90% of links do not cross each other) (Munteanu, 2006). In addition to these criteria there have been some attempts to measure the degree of comparability according to *distribution of topics* and *publication dates* of documents in comparable corpora to estimate the *global comparability of the corpora* (Saralegi et al., 2008). ACCURAT will research criteria of comparability for different document groups with different types of parallelism, e.g., translated texts, texts on the same topic, texts on comparable topics, etc.

As we will focus on under-resourced languages and domains, some of the existing methods for detecting parallel sentences are not always applicable due to the lack of initial resources. For example, a simple word-overlap filter for comparable corpora needs sufficient parallel resources and a number of lexical resources specific to under-resourced languages and narrow domains (e.g., bilingual dictionaries, semantic lexica). ACCURAT research results could be portable to other comparable corpora in under-resourced areas resulting in a language- and domain-independent methodology.

Parallelism on the level of individual sentences will be studied in cases of rough translation equivalents, e.g., when the same event is reported in two different languages, as well as in cases of structural equivalents, e.g., when two conceptually similar events are discussed involving different entities in each language, such as names of organizations, persons, quantities or dates.

4.1.2. Metrics of Comparability and Parallelism

Using defined criteria for parallelism, we would like to develop formal automated metrics for determining the degree of comparability.

Recent studies (Kilgarriff, 2001; Rayson and Garside, 2000) have added a quantitative dimension to the issue of comparability by studying objective measures for detecting how similar (or different) two corpora are in terms of their lexical content. Further studies (Sharoff, 2007) investigated automatic ways for assessing the composition of web corpora in terms of domains and genres. We will study and investigate existing measures and metrics for assessing corpus comparability and document parallelism. Different existing measurement techniques, such as counting word overlap, vector space models (including both bag of words and document structure sensitive approaches), cosine similarity, classification scores, etc. will be explored and combined. The methods of detecting similar documents and sentences in a comparable corpus will be evaluated for precision and recall.

4.2. Methods and Techniques for Building a Comparable Corpus from the Web

Although there are many more potential data sources for comparable corpora than there are for parallel texts, and they are easily accessible via the web, the problem of how to collect these data automatically for under resourced languages and for narrow domains poses a significant technical challenge.

We will begin with building general, i.e. non-domain specific, corpora for under-resourced languages by exploring the limits of techniques that have been developed for extracting parallel corpora for well-resourced languages – for example those exploiting URL and HTML structure, document and text chunk length and basic content matching (Resnik and Smith, 2003; Zhang et al., 2006; Shi et al., 2006). Based on preliminary investigations for the ACCURAT languages, the volume of parallel pages obtainable in this way is too low to yield satisfactory statistical MT models or to extract satisfactory lexical resources on their own. Still such pages, when they exist, are useful for seeding or supplementing lexical resources for use in searching/assembling comparable corpora. Hence we will start by building tools based on existing techniques for automatically building parallel corpora from the web for application to under resourced languages.

Given the paucity of web page pairs that are actual translations for under-resourced languages, we seek pairs of web documents that contain individual sentences which are translations or, weaker still, sentence or phrasal near equivalents. One likely source of such documents is news web sites where one news provider provides news in multiple languages (e.g. Agence France Presse, Xinhua News, Reuters, CNN, BBC). Stories on such sites may not be direct translations, but are likely to share considerable content. Munteanu and Marcu (2005) build their approach to extracting parallel sentences from comparable corpora around such sites, exploiting the LDC gigaword corpora for Chinese, Arabic and English drawn from Agence France Presse and Xinhua news. Unfortunately, none of these major news providers offer services in Croatian, Estonian, Greek, Latvian, Lithuanian, Slovenian or Romanian. However, ACCURAT will explore the underlying idea that contemporaneous news stories in multiple languages will be topically similar by crawling major national monolingual news providers and building comparable corpora of news documents. This will be done by

- restricting the news categories crawled to categories likely to contain stories shared between language communities, e.g. international news, international sporting events (Bekavac et al., 2004);

- restricting date ranges so that documents are likely to be reporting the same events (Munteanu and Marcu, 2005);
- exploring content checking during corpus collection to increase the likelihood that stories are on the same topic, e.g. presence of multiple common named entities.

One question that needs investigation is whether it is better to assemble monolingual corpora independently in multiple languages using the same constraints for each language, or, whether constraints should be established for one language and from the crawled documents meeting these constraints generate queries for other languages using cross-language IR techniques.

Aside from contemporaneous news reports in different languages, another under-exploited source of comparable texts is Wikipedia. There are now a substantial number of Wikipedia articles in each of the under-resourced languages ACCURAT aims to address (Croatian - 65 466, Estonian - 66 308, Greek - 44 173, Latvian 23 058, Lithuanian - 91 315, Slovenian – 79 289, Romanian - 414 091 articles on 24.08.2009). Many of the articles are linked to articles on the same topic in other languages.

ACCURAT will explore the selection of similar documents in multiple languages from Wikipedia. A primary approach is to crawl Wikipedia for comparable articles. The terms and multi-word units that are gathered in this crawl will then be used to seed comparable corpora searches (Bekavac and Tadić, 2008). Here we will be issuing multi-language queries to web search engines to locate such corpora. The similarity of different languages will be tested on different levels, starting from the level of headwords to the level of HTML links that form the structure of relations to other concepts worded as single-word units or multi-word units.

Another approach to generating effective searches will be to issue structured searches, looking for pages written in one language, which are linked to pages written in another language. Also learning the typical tags and text found in links between comparable articles will be examined.

In sum, we propose to explore three classes of techniques to address the problem of automatically assembling comparable corpora for under-resourced languages:

- techniques based on URL and HTML structure, geared at finding web pages which are translations of each other on multilingual sites or which point to related material in other languages
- techniques based on exploiting genre, topicality and shallow content matching to find comparable texts, e.g. news texts in the same category on the same date mentioning the same named entities are likely to report the same events
- techniques based on exploiting cross-language linkages between articles in Wikipedia both to extract comparable corpora directly from Wikipedia and as sources of terms to seed web searches to expand such corpora.

4.3. Techniques for Extraction of Lexical, Terminological and Other Linguistic Data from Comparable Corpora

Multi-level alignment of documents, paragraphs, sentences, phrasal units, named entities and terms for comparable corpora is much more challenging than for parallel corpora.

In parallel corpora, a source language text is translated into one or more sentences in the corresponding target language text and the order of sentences in the two texts tends to be more or less the same. Relatively simple sentence alignment algorithms (e.g., Gale and Church 1991) have proven quite successful at this task and the resulting sentence-aligned texts may then be directly exploited by statistical MT systems.

For comparable corpora the situation is much less straightforward, since, depending on the nature of the comparable corpus, only some or perhaps none of the sentences in any pair of texts from the two languages will be translations of each other. Thus, non-alignment of sentences may well be the norm, and even in cases where two texts communicate information on the same topic (e.g. the same news story), the ordering of information, distribution of information over sentences and the inclusion or exclusion of additional information makes the alignment task extremely challenging.

ACCURAT will address this challenge by investigating a number of multi-level alignment methods for comparable corpora. While our focus and novel contributions are on the alignment of, and acquisition of bilingual lexical resources from comparable corpora, we do not exclude the use of existing parallel corpora. On the contrary, starting from whatever parallel resources are available, we will extract at least seed lexical knowledge to be used in, and enhanced by, the process of aligning comparable corpora.

4.3.1. Selection of Similar Documents from a Comparable Corpus

Given a comparable corpus consisting of documents in two languages, L1 and L2, the first step is to find similar documents in L1 and L2.

Typical approaches involve treating a document in the L1 collection as a query and then using cross-language information retrieval (CLIR) techniques to retrieve the top n documents from the L2 collection (Muntenu and Marcu, 2005, Quirk et al., 2007). This approach requires some sort of bilingual dictionary for use in query translation.

One innovation will be the exploration of bootstrapped bilingual lexical resources: initial bilingual lexicons used for text and sentence alignment will lead to new lexical translation mappings and those with the most confidence will be added to the bilingual lexicons for use in subsequent iterations of text and sentence alignment.

4.3.2. Phrasal Alignment

After similar documents are selected, similar text fragments need to be identified. These fragments may be

sentences or possibly only phrases.

Recent research results have shown that in most cases methods designed for parallel texts perform poorly for comparable corpora. For example, most standard sentence aligners exploit the monotonic increase of the sentence positions in a parallel corpus, which is not observed in comparable corpora.

ACCURAT will investigate how successful the reified sentence aligner (Ceaușu et al. 2006) is in aligning similar sentences in comparable corpora. This reified sentence aligner, based on SVM technology, builds feature structures characterizing a pair of sentences considered for alignment (number of translation equivalents, ratio between their lengths, number of non-lexical tokens, such as dates, numbers, abbreviations, etc., word frequency correlations). These feature structures are afterwards classified as describing GOOD or BAD sentence alignments with respect to experimentally determined thresholds. This aligner has been evaluated and has an excellent F-measure score on parallel corpora, being able to align N-M sentences. It is much better than Vanilla aligner[3] and slightly better than HunAlign[4]. The state-of the-art sentence aligner is Moore's (2002), but this aligner produces only 1-1 alignments (almost perfect), loosing N-M alignments (which downgrades its F-measure score). As comparable corpora do not exhibit the monotonic increase of aligned sentence positions, we anticipate that many of the alignments will be of the type 0-M, N-0 and N-M sentences, thus this alignment ability is a must. The SVM approach to sentence alignment has the advantage that it is fully trainable, the statistical parameters being learnt from the training examples (both positive and negative ones).

Another promising method for identifying similar sentence pairs within comparable corpora, proposed by Munteanu and Marcu (2005), will be also investigated. To select candidate sentences for alignment, they propose a word-overlap filter (half the words of the source language sentence have a translation in the target language sentence) together with a constraint on the ratio of lengths of the two sentences. Given two sentences that meet these criteria, the final determination of whether they are or are not parallel sentences is made by a Maximum Entropy classifier trained over a small parallel corpus, using such features as percentage of words with translations (according to the dictionary), length of sentences, longest connected and unconnected substrings. We will expand this method to sentences / paragraphs which are only to some extent translations of each other, thus adapting the proposed method to comparable corpora.

A challenging research avenue for detecting meaning-equivalent sentence pairs within comparable corpora is using cross-lingual Q&A techniques. The main idea is to exploit dependency linking (Ion and Tufis, 2007) and the concepts of superlinks and chained links (Irimia, 2009) for determining the most relevant search criteria. The keywords, extracted from the dependency linking of a source paragraph/sentence, will be translated (using whatever bilingual resources available, e.g. aligned wordnets, terminology resources or bilingual lexicons - where available, seed translation-pair lists extracted from existing parallel corpora) into a target language and available search engines will look for the most relevant candidate paragraph/sentences. The possible pairs of translation equivalent textual units will be scored by a reified sentence aligner and will be accepted or rejected based on previously determined thresholds.

[3] http://nl.ijs.si/telri/Vanilla

[4] http://mokk.bme.hu/resources/hunalign

4.3.3. Named Entity and Terminology Alignment and Extraction

Finding common named entities (NEs) or technical terms in phrases from texts in different languages is a powerful indicator that the phrases may be translation equivalents, and their absence almost certainly suggests that the phrases are not equivalents (modulo anaphora).

Named entities and many technical terms are typically not found in general purpose lexicons and so their mapping must be established in other ways. Such multi-word expressions typically fall into two types: those which are more or less phonetically equivalent in two languages (e.g. person names like "Barack Obama" – "Barack/Baraks Obama" in Croatian/Latvian and biological terms like "photosynthesis" – "fotosintēze" in Latvian) and those some or all of whose component words are translated individually (e.g. "Black Sea" – "Melnā jūra" in Latvian). In cases where the NEs or terms are not phonologically related, i.e. contain component words that are translations of each other, entity type equivalence together with dictionary matching on component words may be used to align them. In cases where they are phonologically related, however, a process of matching based on transliteration similarity may be used. It is well known that even NE's that are phonologically equivalent across languages are frequently not orthographically equivalent thus to perform named entity matching requires transliteration from the writing system of one language to that of another.

Transliteration can be performed either orthographically or phonetically. In orthographic approaches (e.g. Aswani and Gaizauskas, 2005) possible cross-language n-gram character mappings observed in training data can be recorded and then, for test names, candidate sequence transliterations can be proposed and scored against the candidate name equivalent using string similarity measures, such as edit distance. In phonetic-based approaches (e.g. Kondrak, 2000; Mani et al., 2008), names are transduced into a phonetic representation and then candidate matches are determined using edit distance measures with learned thresholds.

In the ACCURAT project we will explore both approaches, developing adaptive HMM and/or CRF-based techniques (e.g. Zhou et al., 2008) trained on name pairs gathered initially from parallel training data and then bootstrapped using lexicons derived in the project.

We will also exploit new advances in adaptive, semi-supervised NE recognition (e.g. Nadeau, 2007) that allow powerful NERC systems to be built for a wide range of entity types from only a handful of examples in each entity class together with suitable corpora. These techniques have not been extensively explored for languages other than English.

Since terminology is of utmost importance in the translation of technical documents, automated terminology extraction will be a basic facility for development of MT systems for narrow domains. Our work will be focused on exploring the use of existing term extraction techniques for terms within the narrow domains. Various techniques exist for identifying terms within a domain-specific (monolingual) corpus and we will build on these. One of techniques is supervised and weakly supervised for semantic labeling of terms within specialist domains in English (e.g. biomedicine; Roberts et al, 2008) which should be relatively portable across languages. Other techniques do not attempt semantic labeling but just attempt to recognize multiword units that are domain specific terms (Bourigault et al, 2000).

Research on bilingual terminology extraction has started recently and relies on assumption that words with same meaning in different languages tend to appear in the same context (Rapp, 1995). The most common approach is to use context vectors and evaluate candidate translations. On single words this approach demonstrated good results (e.g. Chiao and Zweigenbaum, 2002). Recently Daille and Morin (2008) adapted this direct context vector approach for single and multi-word terms and added compositional translation methods for French-Japanese languages. This method increases by 10% the results of Morin et al. (2007), however they are still rather low for multi-word terms.

4.3.4. Relation Extraction for Phrasal Alignment

Another novel way information extraction techniques can assist in aligning comparable corpora is through the identification of cross-language mappings between relation-expressing contexts. Hasegawa et al. (2004) propose a technique for unsupervised relation discovery in texts, whereby contexts surrounding pairs of NEs of given types are extracted and then clustered, the clusters correspond to particular relations (e.g. the relation "company X ACQUIRES company Y" may be expressed as "X's purchase of Y", "X has agreed to buy Y"). This technique achieves impressive results and could be used to align relation expressing contexts as follows. First relation clusters could be established monolingually, given NERC tools in each language. These clusters could then be aligned cross-lingually, using aligned sentence pairs containing NE pairs found in the clusters, aligned sentences coming either from a small amount of parallel data or from high confidence alignments in the comparable corpus. Once relation clusters are aligned cross-lingually, presence of a pair of NEs from an aligned relation cluster in an L1 and L2 sentence pair would constitute evidence that sentences should be aligned.

4.4. Comparable Corpora in Machine Translation Systems

To evaluate the efficiency and usability of the approach proposed in the ACCURAT project for under-resourced areas of MT, we will integrate research results into SMT and rule-based systems. We will measure improvements from applying acquired data against baseline results from SMT and RBMT systems and will evaluate the ACCURAT project results in practical applications. The ways how comparable corpora will be integrated and evaluated in MT are described in Eisele and Xu (2010).

5. Conclusions

The ACCURAT project has the ambitious goal of developing solutions for the application of comparable corpora in machine translation. Previous research and the planned approach described in this paper allow us to expect promising results from this research.

6. Acknowledgements

The project has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n° 248347. Part of this work was funded by European Social Fund.

Many thanks for project preparation to colleagues in ACCURAT partner organizations: Andreas Eisele from DFKI (Germany), Serge Sharoff from University of Leeds (UK), Gregor Thurmair from Linguatec (Germany), Nikos Glaros from Institute for Language and Speech Processing (Greece), Marko Tadić from University of Zagreb (Croatia), Boštjan Špetič from Zemanta (Slovenia).

7. References

Alegria, I., Ezeiza, N., Fernandez, I. (2008). Translating Named Entities using Comparable Corpora. In *Proceedings of the Workshop on Comparable Corpora, LREC'08*, pp. 11-17.

Aswani, N., Gaizauskas, R. (2005). Aligning Words in English-Hindi Parallel Corpora. In *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pp. 115-118.

Bekavac, B., Osenova, P., Simov, K., Tadić, M. (2004). Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croatian. In *Proceedings of the 4th Language Resources and Evaluation Conference: LREC04*, Lisbon, pp. 1187-1190.

Bekavac B. and Tadić M. (2008). A Generic Method for Multi Word Extraction from Wikipedia. In *Proceedings of ITI2008 Conference*, SRCE, Zagreb, pp. 663-667. Chiang D. (2007) Hierarchical Phrase-Based Translation. In *Computational Linguistics* 33(2): 201-228.

Bourigault, D., Jacquemin, C. and L'Homme, M. (eds.). (2002), *Recent Advances in Computational Terminology*, CNRS, ERSS, Université Toulouse-le-Mirail / CNRS-LIMSI, Orsay, France /

Université de Montréal.

Ceauşu A., Ştefănescu D., Tufiş D. (2006). Acquis Communautaire Sentence Alignment using Support Vector Machines. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 2134-2137*.

Chiang, D. (2007) Hierarchical Phrase-Based Translation. In *Computational Linguistics* 33(2): 201-228.

Chiao, Y., Zweigenbaum, P. (2002) Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In : *COLING 2002*, pp. 1208-1212, Tapei, Taiwan.

Daille, B. and Morin E. (2008) An Effective Compositional Model for Lexical Alignment. In : *Proceedings of IJCNLP-08*, pp. 95-102.

Daille, B., Morin, E. (2005) French-English Terminology Extraction from Comparable Corpora. In : *IJCNLP 2005*, pp. 707-718

EAGLES. (1996). Preliminary recommendations on corpus typology. Electronic resource: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., Yu Chen (2008) Hybrid Machine Translation Architectures within and beyond the EuroMatrix project. In *Proceedings of EAMT*, Hamburg.

Eisele, A. and Xu, J.(2010). Improving machine translation performance using comparable corpora. In: *Proceedings of 3rd Workshop on Building and Using Comparable Corpora* , Malta.

Gale, W. and Church, K. (1991). A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 177-184.

Goutte, C., Cancedda, N., Dymetman, M., Foster, G. (eds.). (2009). *Learning Machine Translation*. The MIT Press. Cambridge, Massachusetts, London, England.

Hasegawa, T., Sekine, S., and Grishman, R. 2004. Discovering relations among named entities from large corpora. In *ACL '04*.

Hewavitharana, S. and Vogel, S. (2008). Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources. In *Proceedings of the Workshop on Comparable Corpora, LREC'08*, pp. 7-10.

Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceedings of the Association for Computational Linguistics*, pp. 236-243.

Ion, R. and Tufiş, D. (2007). RACAI: Meaning Affinity Models. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007*), pp. 282-287, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Irimia, E. (2009). Metode de traducere automată prin analogie. Aplicaţii pentru limbile română şi engleză. (Methods for Analogy-based Machine Trannslation. Aplications for Romanian and English). PhD thesis, March 2009.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

Koehn, P., Birch, A., Steinberger, R. (2009).462 machine translation systems for Europe. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, August 26-30, 2009, Ottawa, Ontario, Canada; pp. 65-72.

Koehn, P. and Hoang, H. (2007). Factored Translation Models. In *Proceedings of EMNLP'07*.

Koehn, P., Och, F. J., Marcu, D. (2003). Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*.

Kondrak, G. (2000). A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics ANLP-NAACL'00*, pp. 288-295.

Mani, I., Yeh, A., Condon, S. (2008). Learning to Match Names Across Languages. In *Proceedings of the COLING 2008 Workshop on Multi-source Multilingual Information Extraction and Summarization MMIES'08*, pp 2-9.

McEnery, A.M. and Xiao, R.Z. (2007). Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist.* Translating Europe. Multilingual Matters, Clevedon, UK.

Maia, B. (2003). What are Comparable Corpora? *Electronic resource*: http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc

Maia, B. and Matos, S. (2008). Corpógrafo V.4 – Tools for Researchers and Teachers Using Comparable Corpora. In *Proceedings of the Workshop on Comparable Corpora, LREC'08*, pp. 79-82.

May, J. and Knight, K. (2007). Syntactic Re-Alignment Models for Machine Translation. In *Proceedings of EMNLP-CoNLL'07*.

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Machine Translation: From Research to Real Users* (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, pp. 135—244.

Morin, E., Daille, B., Takeuchi, K., Kageura, K. (2007) Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In: *ACL 2007*

Munteanu, D. (2006). Exploiting Comparable Corpora (for automatic creation of parallel corpora). *Online presentation. Electronic resource*: http://content.digitalwell.washington.edu/msr/external_release_talks_12_05_2005/14008/lecture.htm

Munteanu, D. and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel

Corpora. *Computational Linguistics*, 31(4): 477-504.

Munteanu, D., Fraser, A., Marcu, D. (2004). Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT / NAACL'04.*

Nadeau, D. (2007). Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision. PhD Thesis, University of Ottawa, 2007.

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

Quirk, C., Udupa, R., Menezes, A. (2007). Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction. In *Proceedings of MT Summit XI, European Association for Machine Translation.*

Rapp, R. (1995). Identifying Word Translations in Non-Parallel Texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Comparing Corpora Workshop at ACL'00*, pp. 1-6.

Resnik, P. and Smith, N. (2003). The Web as a Parallel Corpus. *Computational Linguistics*, 29(3) pp. 349-380.

Roberts, A., Gaizauskas, R., Hepple, M., Guo, Y. (2008). Combining terminology resources and statistical methods for entity recognition: an evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008).

Saralegi, X., San Vicente, I., Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the Workshop on Comparable Corpora*, *LREC'08*.

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. In *Proceedings of Web as Corpus Workshop*. Louvain-la-Neuve.

Shi, L., Nie, C., Zhou, M., Gao, J. (2006). A dom tree alignment model for mining parallel data from the web. In *Joint Proceedings of the Association for Computational Linguistics and the International Conference on Computational Linguistics*, Sydney, Australia.

Sinclair J. (1987) (eds.) Looking up: an account of the COBUILD Project in lexical computing. Collins, London and Glasgow.

Skadiņa, I., Brālītis, E. (2009). English-Latvian SMT: knowledge or data? In: *Proceedings of the 17th Nordic Conference on Computational Linguistics NODALIDA.* NEALT Proceedings Series, Vol. 4, pp. 242–245.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006) The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC'06*.

Thurmair, G. (2006). Using Corpus Information to Improve MT Quality. In *Proceedings of the Workshop LR4Trans-III, LREC*, Genova.

Tillmann, C. (2003). A projection extension algorithm for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural language processing*.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, pp. 153-160, Trento, Italy, April 2006.

Utiyama, M., Isahara, H. (2003). Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 7-12 July 2003, Sapporo, Japan.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., Waibel, A. (2003). The CMU Statistical Machine Translation System. In *Proceedings of MT-Summit IX*.

Yamada, K. and Knight, K. (2001). A Syntax-Based Statistical Translation Model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 523–530, Toulouse, France, July.

Yamada, K. and Knight, K. (2002). A Decoder for Syntax-Based Statistical MT. In *Proceedings of the Conference of the Association for Computational Linguistics, ACL'02*.

Zhang, Y., Wu, K., Gao, J., Vines, P. (2006). Automatic Acquisition of Chinese-English Parallel Corpus from the Web. In *Proceedings of 28th European Conference on Information Retrieval*.

Zhao, B., Vogel, S. (2002). Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, p.745, December 09-12, 2002.

Zhou, Y., Huang, F., Chen, H. (2008). Combining probability models and web mining models: a framework for proper name transliteration. In *Information Technology and Management* 9(2).