Languages for Special Purposes in a Multilingual, Transcultural World

Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria

http://lsp2013.univie.ac.at/proceedings

Translation technology for terminology in higher education

Mirela-Ştefania Duma; Melania Duma; Walther v. Hahn; Cristina Vertan

Cite as: Duma, M-Ş. et al. (2014). Translation technology for terminology in higher education. In G. Budin & V. Lušicky (eds.), *Languages for Special Purposes in a Multilingual, Transcultural World, Proceedings of the 19th European Symposium on Languages for Special Purposes, 8-10 July 2013, Vienna, Austria*. Vienna: University of Vienna, 220-227.

Publication date: July 2014

ISBN: **978-3-200-03674-1**

Translation technology for terminology in higher education

Mirela-Ştefania Duma; Melania Duma
NATS Division, Faculty of Mathematics, Informatics and Natural Sciences, University of Hamburg
Germany

Walther v. Hahn; Cristina Vertan
Research Group "Computerphilologie", University of Hamburg
Germany

Correspondence to: vertan@informatik.uni-hamburg.de

Abstract. The Bologna process together with the EU enlargement brought new stimuli to the student mobility all over Europe. Even though it is generally accepted that English is used for communication as a lingua franca, in most cases the courses are delivered in the language of the host university. It is recommended for exchange students to provide a B1 language certificate, but this is not sufficient for understanding specialized vocabulary. For translating domain specific terms a student has to use either dedicated dictionaries (available only for limited domains and language pairs) or freely available machine translation (MT) systems. Current freely available MT-systems rely on large amounts of training data, thus the quality of the translation is highly dependent on the similarity between the input and the training material. The aim of this paper is twofold: to analyse the behaviour of three online available systems exposed to excerpts from curriculum descriptions and to present an approach for domain adaptation for open domain MT, discuss its improvements as well as its limitations. We focused our study on the German-English and German-Romanian language pairs. We perform both automatic evaluation and human evaluation and analyse the degree of correlation between them.

Keywords. Specialised texts, open domain machine translation, domain adaptation, terminology translation.

1. Introduction

The adoption of the Bologna treaty by most part of the universities within the enlarged European Union brought new impulses to the student mobility across all 27 countries. Even though it is generally accepted, that English is used as a lingua-franca for daily communication by most exchange students, in most cases the courses are delivered in the language of the host university (especially at the bachelor level). Exchange students are encouraged to provide a B1-level certificate for the language of the host university, however this is not enough for understanding specialized vocabulary. The preparation of the exchange (i.e. the choice of courses for the "Learning Agreement") is also difficult as students do not understand completely the course descriptions, requirements etc.

For universities the variety of languages spoken by the exchange students, as well as the dynamic character of the curriculum descriptions, is a real challenge. Given this setting (many language pairs, dynamic texts and specialised domains) human translation cannot be a solution. Specialised lexicons are rare and predominantly available from/into English, German or Spanish.

Recent developments in machine translation make this technology a possible solution for curriculum translation. However, current MT-technology strongly relies on large training material, thus using a translation system in an open domain setting (like university curricula) poses several challenges and eventually leads to the implementation of domain adaptation strategies.

It is frequently assumed that the quality of on-line translation engines is decreasing when used with specialised texts, but only few systematic analyses were done, especially with less-resourced language pairs (e.g. Romanian-German).

In this paper we present the analysis of three on-line available systems, one of them being developed from scratch and thus giving the opportunity to embed domain-adapted translation models. We will use the term *out-of-domain* to refer to the general domain used for training the system (its source domain). The term *in-domain* is used to refer to the specific domain, the domain used for testing (target domain). In section 3 we introduce the three translation engines tested and discuss the domain adaptation strategy. Section 3 presents the results of the automatic and manual evaluation and section 4 is reserved for conclusions and further work.

2. Machine translation engines

We decided to use three on-line translation engines, two of them only available as web services and therefore working as black-boxes: Google Translate[1] and Bing[2]. Both systems are based on corpus-based machine-translation paradigms, but neither the training data nor the translation models and best-candidate selection strategies are known.

The third machine translation engine was developed within the EU funded project ATLAS[3] (Applied Technology for Language Enhanced CMS)

The ATLAS system basically is a generic framework for web content management, which makes use of state-of-the art text technologies for information extraction and to cluster documents according to a given hierarchy. A text summarization module and a machine translation engine as well as a cross-lingual semantic search engine are embedded.

The system for the moment is handling six languages (Bulgarian, English, German, Greek, Polish and Romanian) from four language families. However, the chosen framework allows additions of other languages at a later point of time.

The core online service of the ATLAS platform is "i-Publisher", a powerful web-based instrument for creating, running and managing content-driven web sites. It integrates the language-based technology to improve content navigation e.g. by interlinking documents based on extracted phrases, words and names, providing short summaries and suggested categorization concepts. Currently two different thematic content-driven web sites are being built on top of the ATLAS platform, "i-Librarian" and "EUDocLib", both using i-Publisher as content management layer. iLibrarian is a user-oriented web platform which allows users to maintain a personal workspace for storing, sharing and publishing various types of documents and have them automatically categorized into appropriate subject categories, summarized and annotated with important words, phrases and names, as well as translated. EUDocLib is a publicly accessible repository of EU legal documents from the EUR-LEX collection with enhanced navigation and multilingual access techniques.

A key component of the ATLAS system is the machine translation engine. Its development was particularly challenging as the system is open-domain and has to handle different text-genres. Additionally, the considered language-pairs belong to the so called *less resourced* group, for which bilingual training and test material is available only to a limited amount (Vertan 2012).

The machine translation engine is integrated in two distinct ways into the ATLAS platform:

- for i-Publisher Services (generic platform for generating websites) the MT is serving as a translation aid for publishing multilingual content. Text is submitted to the translation engine and the result is subject to human post processing,
- for i-Librarian and EuDocLib (web services for collecting documents) the MT-engine provides a translation for assimilation, which means that the user retrieving documents in different languages will use the engine in order to get a clue about the documents, and decide if he will store them. If the translation is considered as acceptable it will be stored into a database.

Being supposed to work in an open-domain setting the main challenge for the MT-engine is

to handle specialised data. In contrast with other on-line sytems, the ATLAS-engine has the advantage of accessing linguistic processing chains integrated within the ATLAS content managemnt systems.

We decided to implement a user-proactive strategy as follows: each domain uploaded to the system is first assigned automatically to a node in a classification hierarchy. With the prototype ATLAS system we provided a hierachy following the Library of Congress Classification[4]. From this large classification we selected thirteen nodes to which we attached translation-information vector-labels. A translation-information vector-label has fifteen components (the number of translation pairs); each vector component corressponds to a langauge pair and informs if a specific trained model for that domain is available.

A general translation model was trained on the JRC-Acquis corpus[5]. The JRC-Acquis corpus (Steinberger et. al. 2006) is a multilingual parallel corpus for 22 European languages consisting of paragraph alignments for 231 pairs[6] of languages. The data is made up of a selection of european documents refered to as Acquis. This term identifies the body of common rights and obligations that bind all the member states from the European Union. The choice of using this corpus is motivated by the fact that it is the only free –available and size –relevant ressource available for all languages involved in the system.

The in-domain corpora were colected partially manually, partially were generated automatically using English as a pivot language. They are all small corpora having between two and five thousand sentence-pairs.

For the analysis in this paper we selected the domains Mathematics and Biology. The choice is motivated as follows:

- Mathematics is a very technical domain completely different from the general JRC-Acquis corpus, both in syntax and lexical coverage
- Biology is very different in terms of lexical coverage from the general domain

The language pairs selected for the experiments were German-Romanian and German-English.

2.1. ATLAS statistical machine translation component

The ATLAS translation engine uses a hybrid approach combining an example-based component with a statistical-based one. The example-based component is working only when parts of the input are retrieved identical in the translation database, otherwise the statistical component is working. Thus we will describe here in detail just the statistical component.

Based on the large out-of-domain corpus JRC-Acquis, a translation model is trained which gives the probability that a sequence of words in the target language is the translation of another sequence of words in the source language. The translation probability is computed as in (1):

$$(1) \quad P(t \mid s) = \frac{1}{Z} exp\left(\sum_{i=1}^{M} \lambda_i h_i(t,s)\right)$$

where t is a phrase in the target language, s is a phrase in the source language, Z is a normalization factor that ensures that the result is between 0 and 1, M is the number of features that the translation system has, λ_i is a corresponding weight for the feature function $h_i(t,s)$. A baseline translation system includes the following feature functions:

- $P(t \mid s)$ and $P(s \mid t)$: the phrase probabilities in both directions
- $P_l(t \mid s)$ and $P_l(s \mid t)$: the lexical probabilities in both directions which show how well individual word translates to each other

- *P(t)*: the language model which tells how likely a candidate translation is fluent in the target language
- *d(s,t)*: the distortion model which reorders phrases
- *W(t)*: word penalty which penalizes very long or short target sentences.

2.1.1. Domain adaptation strategy

Domain adaptation became a major research field in machine translation during the last years. Many heuristics were proposed, however they are highly dependent on the particular context for which they were developed. An attempt to classify domain adaptation strategies can be found in (Duma and Vertan, 2013). Based on this classification we can summarize a generic domain adaptation work-flow for SMT, as in Fig. 1. Here the focus is on the model approach. The training data consists of out-of-domain used to infer a SMT baseline model and in-domain data which is used together with the inferred model to adapt it. The testing data belongs to the same domain as the in-domain data.

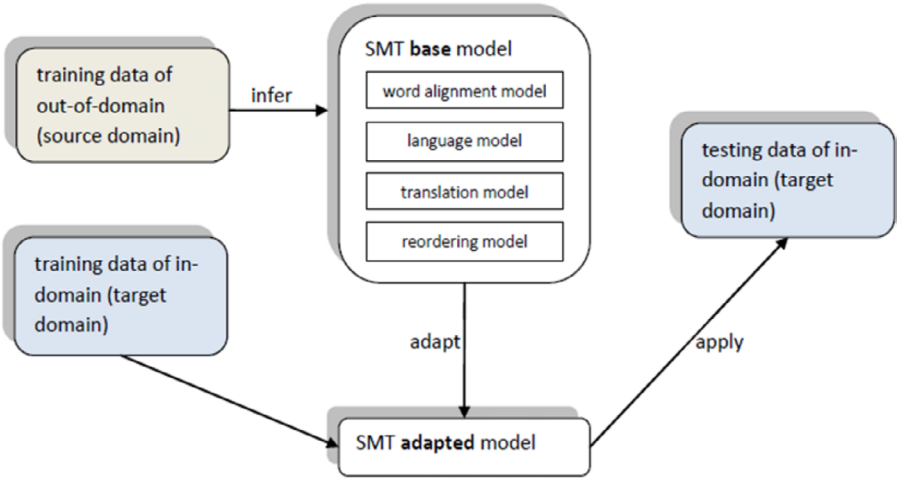

Figure 1: Domain adaptation setup for SMT (Duma and Vertan, 2013)

2.1.2. Selected state of the art method for domain adaptation

The baseline translation systems and the adapted systems incorporated into the ATLAS-engine were built with the Moses[7] framework.

We chose a domain adaptation method based on linear interpolation as in (Koehn and Schroeder, 2007). The translation probability is computed as in (2)

$$P(t|s) = wP_{in}(t|s) + (1-w)P_{out}(t|s) \quad (2)$$

where

- $P(t|s)$ is the conditional probability of a phrase-pair,
- $P_{in}(t|s)$ is the conditional probability of a phrase-pair corresponding to the in-domain and
- $P_{out}(t|s)$ is the conditional probability of a phrase-pair corresponding to the out-of-domain.
- w is the interpolation weight, $0 \leq w \leq 1$.

Using the SRILM language model toolkit[8] (Stolcke 2012), we compute in this order:

- the language model for the in-domain,
- the language model for the out-domain,
- and the weight

Consequently by means of (2) an interpolated language model is created.

3. Evaluation and analysis of the results

For the test set we used 100 sentences from each in-domain corpus. The automatic evaluation consisted of using the BLEU metric (Papineni et. al. 2002) as an evaluation metric. BLEU counts identical n-grams in MT-output and reference translation. For the manual analysis, we extracted terms and multi-word expressions specific to the in-domain and compared the translation from two online translation systems, Google and Bing and the machine translation engine of the ATLAS system.

3.1. Automatic evaluation

In Tab. 1, the BLEU scores are given for the baseline ATLAS-Engine (SMT without domain adaptation), for the adapted ATLAS-engine system, for Google and for Bing. At a first sight BLEU scores for Google and Bing are better than the other ones. These systems do produce always fluent output. It can be observed that the adapted systems gave a much better BLEU result than the baseline systems, indicating that the language model interpolation method is a good approach inducing an increase of the performance even with small in-domain available training data.

BIOLOGY		
	DE-EN	DE-RO
Baseline system	15.04	4.69
Adapted system	21.09	10.19
Google	54.16	28.81
Bing	23.23	19.45
MATHEMATICS		
	DE-EN	DE-RO
Baseline system	20.86	4.68
Adapted system	28.88	10.94
Google	56.97	36.87
Bing	43.76	26.73

Table 1: BLEU scores for Baseline and Domain Adapted ATLAS engine, Google and Bing for Mathematics and Biology

3.2. Manual analysis

For the manual analysis we selected manually from the 100 sentences of the test sets terminological expressions characteristic to the Mathematics domain, and observed their translation by the three systems. This comparison is presented in Tab. 2 where “oov” means “out of vocabulary words”.

At a first glance we observe the relative high number of out of vocabulary words among the output presented by the ATLAS -engine. This is due to the small in-domain training data for the given language pair. We observe however a quite high number of terminological expressions which are translated at least as good as Google and Bing by the ATLAS adapted system. Overall the adapted system does not present cases of wrong semantic translation, which leads to the idea that with a larger in-domain training data it could overpass the quality of the other on-line translation engines.

Word or construction	Reference	Adapted System ATLAS	Google Translate	Bing
Logischen Schließens	a deduce logic	oov	deductie logica	deducerea logică
Idealisierte Denkmodelle	modele idealizate ale gândirii	oov	cu modele idealizate de gândire	cu modele idealizat
Zeichenketten	şiruri de caractere	oov	siruri de caractere	siruri de caractere
ideale Gedankenkonstruktionen	rationamente abstracte	ideal + oov	gânduri în domeniul construcțiilor ideale	constructii ideală de gânduri
innermathematischen	intra-matematice	oov	intra-matematice	intra-mathematical
Seitenlänge	lungimea laturii	**lungimea laturii**	lungimea partea sa	partea lui
geometrische Figur in der Ebene	figură geometrică în plan	*geometrische figur* în de plan	formei geometrice în plan	un poligon în planul
Strecken	linii	distante	linii	linii
Ebene	nivelul	**nivel**	nivelul	nivel
Quadratdiagonalen	diagonalelor patrati	oov	diagonalelor pătrat	de pătrat diagonala
Quadratwurzel	rădăcina pătrată	**rădăcina pătrată**	rădăcina pătrată	rădăcina pătrată + piața rădăcini
Satz von Pythagoras	teorema lui Pitagora	coeficientul de Pythagoras	teorema lui Pitagora	teorema lui Pitagora
Längen	lungimi	oov	lungimi	lungimi
Winkeln	unghiuri	**unghiuri**	unghiuri	unghiuri
Abstäde	distanţe	**distante**	distanţe	distanţe
Dreiecke	triunghiuri	oov	triunghiuri	triunghiuri
Quadrate	pătrate	**pătrate**	pătrate	pătrate
Kreise	cercuri	**cercuri**	cercuri	cercuri
Gleichungen	ecuaţii	**ecuaţii**	ecuaţii	ecuaţii
Funktionen	funcţii	funcţii	funcţii	funcţii
Differenzieren	diferenţiere	**diferentiere**	diferenţiere	diferenţiere
Integrieren	integrare	**integrare**	integrare	integrare
Schnittpunkt	Intersecţia	**intersectia**	Intersecţia	Intersecţia
Geraden	linii	**linie**	linii	linii
aufeinander normal stehende Geraden	în mod normal, legate unele de altele ca linie dreaptă		ca de obicei unul de altul în picioare drept	ca reciproc în picioare drepte normale
Dezimalldarstellung	zecimale	**reprezentare zecimal**	zecimale	reprezentarea zecimală
Dezimallstellen	cifre zecimale	mult mai zecimale	cifrele mai zecimale	mai multe zecimale
Trigonometrisches Problem	problemă trigonometrice	oov	problemă trigonometrice	problemă trigonometrice
Trigonometrie	trigonometrie	**trigonometrie**	trigonometrie	trigonometrie
geradlinig	drept	oov	drept	drept

Table 2: Comparative translations for mathematical terminology translation engines

4. Conclusions and future work

In this paper we presented an experiment of using online machine translation systems for translating specialised texts as in curriculum description in higher education. The main reason behind this experiment was to asses to which extent such systems can be used for assisting exchange students in their studies. We conclude that linear interpolation used for domain adaptation for statistical systems following the Moses framework, might out or equally performs third-party on-line systems, assuming that enough in-domain training data is available. Otherwise the out of vocabulary words may decrease the performance of the system.

Another advantage of having the own adapted SMT-system is the complete control on the translation system: one can observe the lacks in language coverage, increase the training data, retrain the model, and eventually involve external linguistic sources. This is not possible with black-box solutions as the other two engines tested in our experiment.

Further work consists in extending the analysis on other domains, retrain the models with larger data as well as investigating other methods for domain adaptation.

5. Notes

[1] http://translate.google.com/

[2] http://www.bing.com/

[3] http://www.atlasproject.eu

[4] http://www.loc.gov/catdir/cpso/lcco/

[5] http://optima.jrc.it/Acquis/index_2.2.html

[6] http://langtech.jrc.ec.europa.eu/Documents/070622_Poster_JRC-Acquis.pdf

[7] http://www.statmt.org/moses/

[8] http://www.speech.sri.com/projects/srilm/

6. References

Cronin, Blaise (1981). The Need For A Theory Of Citing. *Journal of Documentation*, *37*(1), 16-24.

Duma, Mirela-Ştefania & Cristina, Vertan (2013). Integration of Machine Translation in On-line Multilingual Applications - Domain Adaptation. *Translation: Computation, Corpora, Cognition* [Online], Vol. 3, No.1.

Freedman, Aviva, & Adam, Christine (2000). Write where you are: Situating learning to write in university and workplace settings. In Patrick Dias & Anthony Paré (eds.), *Transitions: Writing in academic and workplace settings*. Cresskill, NJ: Hampton Press, 31-60.

Koehn, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Beroldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine; Zens, Richard; Dyer, Chris; Bojar, Ondrej; Constantin, Alexandra & Herbst, Evan (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL),* demonstration session. Prague, Czech Republic.

Koehn, Philipp & Schroeder, Josh. (2007). Experiments in Domain Adaptation for Statistical Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation.* Prague, Czech Republic, 224-227.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*. 311-318.

Pears, Richard & Shields, Graham J. (2010). *Cite them right: the essential referencing guide* (8th ed.). Basingstoke: Palgrave Macmill.

Schroeder, Josh & Koehn, Philipp (2007). The University of Edinburgh System Description for IWSLT 2007. *Proceedings of the International Workshop on Spoken Language Translation.* Trento, Italy, 224-227.

Steinberger, Ralf; Pouliquen, Bruno; Widiger, Anna; Ignat, Camelia; Erjavec, Tomaž; Tufiş, Dan; Varga, Dániel (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *Proceedings of the conference on Language Resources and Evaluation*, 2142-2147.

Stolcke, Andreas (2002). SRILM - An Extensible Language Modeling Toolkit. *Proceedings of International Conference on Spoken Language Processing*. Denver, Colorado.

Vertan, Cristina (2012). Embedding Machine Translatsion in ATLAS Content Management System. *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT) 2012*. Trento, Italy.