Toponym Disambiguation in an English-Lithuanian SMT System with Spatial Knowledge

Raivis Skadiņš
Tilde SIA
Vienibas gatve 75a,
Riga, Latvia
raiviss@tilde.lv

Tatiana Gornostay
Tilde SIA
Vienibas gatve 75a,
Riga, Latvia
Tatiana.Gornostay@tilde.lv

Valters Šics
Tilde SIA
Vienibas gatve 75a,
Riga, Latvia
Valters.Sics@tilde.lv

Abstract

This paper presents an innovative research resulting in the English-Lithuanian statistical factored phrase-based machine translation system with a spatial ontology. The system is based on the Moses toolkit and is enriched with semantic knowledge inferred from the spatial ontology. The ontology was developed on the basis of the GeoNames database (more than 15 000 toponyms), implemented in the web ontology language (OWL), and integrated into the machine translation process. Spatial knowledge was added as an additional factor in the statistical translation model and used for toponym disambiguation during machine translation. The implemented machine translation approach was evaluated against the baseline system without spatial knowledge. A multifaceted evaluation strategy including automatic metrics, human evaluation and linguistic analysis, was implemented to perform evaluation experiments. The results of the evaluation have shown a slight improvement in the output quality of machine translation with spatial knowledge.

1 Introduction and Background

During recent decades the corpus-based strategy has become dominant for machine translation, as it has proven to be more effective both from the point of view of time and labour resources and the quality of the output. The statistical approach has occupied the leading position with the first research results performed in the late 1980s. Since then statistical machine translation (SMT) has become the major focus for many research efforts due to its cost effectiveness doubled with the availability of such open source tools as GIZA++ (Och and Ney, 2003) and Moses (Koehn et al., 2007), as well as parallel text resources on the Internet.

Pure SMT methods (Brown et al., 1993; Koehn et al., 2003) do not use any linguistic knowledge (e.g. morphological information). As a result, they perform better for analytical languages, such as English, with little inflection. Although English and Lithuanian are Indo-European languages and share some grammatical features, they have a wealth of differences. English belongs to the Germanic language group while Lithuanian belongs to the group of Baltic languages. Also, in the morphological typology English is an analytical language in contrast to a synthetic Lithuanian with a rich set of inflections. SMT for synthetic languages with high inflection (e.g. Lithuanian, Latvian, Russian and others) requires larger amounts of training data and additional knowledge to get the same level of performance.

Modern SMT methods use different kinds of additional knowledge (e.g. morphological or syntactical) to build more sophisticated statistical models and improve the output quality of machine translation (see, for example, factored SMT (Koehn et al., 2007), tree-based SMT (Chiang 2007; Marcu et al., 2006; Li et al., 2009); treelet SMT (Quirk et al., 2005). This paper presents an innovative research resulting in an English-Lithuanian statistical factored phrase-based machine translation system based on the Moses toolkit and enriched with semantic knowledge inferred from the spatial ontology.

Using semantic knowledge in rule-based machine translation is not new in the field. In SMT, however, there has been little research in this area[1]. The implemented SMT system that is de-

[1] See, for example, the research on extracting phrasal correspondences that are approximately semantically equivalent for building a full-sentence paraphrasing model that then is applied to a single good reference translation for each sentence in a statistical machine translation development set (Madnani et al., 2008).

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 191–197

scribed in this paper uses semantic knowledge to improve the quality of translation, in particular with regard to the disambiguation of geographical names, or toponyms. Spatial knowledge is added to toponyms in the source text as additional semantic tags, or factors. By adding factors into the source text, the translation accuracy is improved. This is the result of resolving semantic ambiguities in the source language.

The first part of the paper overviews the system design including a description of its functionality and implementation with spatial knowledge. In the second part we focus on the system multifaceted evaluation and its results, as well as potential limitations of the system. Finally, we present conclusions and future plans.

2 System Design

2.1 Functionality

In the overall machine translation theory and in practice English-Lithuanian toponym translation problems have not been researched before. The core functionality of the presented system is a disambiguation of toponyms during the machine translation process. Toponyms are geographical names, or names of places (hydronyms, oronyms, geonyms, oeconyms, etc.). A natural language is ambiguous and toponyms are not exceptions. This fact makes toponyms difficult for processing (e.g. resolution, cross-language information retrieval, human translation and especially machine translation), and due to their linguistic and extra-linguistic nature toponyms require special treatment (Gornostay and Skadiņa, 2009).

There are cases when real-world geographical knowledge is required for the resolution of ambiguous toponyms. The implemented SMT system deals with two types of ambiguity (see Leidner (2007) for the description of possible types of toponym ambiguity). The first type is a referential ambiguity, where a toponym may refer to more than one location of the same type, for example:

- *Georgia* as the US state and the country in Caucasus (English);
- *Riga* as the populated place and the capital of Latvia and as the populated place in the USA, state Michigan (Latvian);
- *Šveicarija* as the village in Lithuania and as the country in Europe (Lithuanian).

The second type of ambiguity is a feature type ambiguity, where a toponym may refer to more than one place of a different type, for example:

- *Tanfield* refers to the populated place as well as the castle in the United Kingdom (English);
- *Gauja* refers to the populated place as well as the river in Latvia (Latvian);
- *Šventoji* as the town near the Baltic Sea as well as the name of 3 different rivers in Lithuania (Lithuanian).

In the implemented system the two described types of toponym ambiguity are resolved using semantic knowledge inferred from the spatial ontology.

2.2 Baseline SMT System

The baseline system was a statistical phrase-based machine translation system based on the Moses toolkit and trained on the following publicly available and proprietary corpora:

- DGT-TM parallel corpus[2] – a publicly available collection of legislative texts in 22 languages of the European Union;
- OPUS parallel corpus – a publicly available collection of texts from the web in different domains[3] (Tiedemann, 2004; Tiedemann, 2009).
- Localization parallel corpus obtained from translation memories that have been created during the localization of software, user manuals and helps.

We also included word and phrase translations from bilingual dictionaries and term translations from EuroTermBank[4] to increase word coverage.

Monolingual corpora for the training of language models were prepared from corresponding monolingual parts of parallel corpora, as well as Lithuanian news articles collected from the web. Bilingual and monolingual resources prepared and used for the baseline SMT system development are represented in Table 1.

Monolingual corpus	**Units**
Lithuanian side of parallel corpora	~4,04 mil.

[2] http://langtech.jrc.it/DGT-TM.html
[3] We chose the EMEA (medical domain) and KDE4 (IT domain) sentence-aligned corpora.
[4] www.eurotermbank.com

Web news	~5,22 mil.
Total	**~9,26 mil.** (filtered)
Bilingual corpus	**Parallel units**
Localization TM	~5,21 mil.
DGT-TM	~1,08 mil.
OPUS EMEA	~1,04 mil.
Dictionary data	~0,27 mil.
EuroTermBank data	~0,1 mil.
KDE4	~0,05 mil.
Fiction	~0,01 mil.
Total (used for the baseline system)	**~7,76 mil.** (filtered)

Table 1. Training corpora.

2.3 Spatial Ontology

The spatial ontology to be integrated into the machine translation process was developed using the ontology language, designed and implemented in the web ontology language (OWL) using RCC-8 properties (Region Connection Calculus) (Randell et al., 1992), and tools developed in the SOLIM project[5]. RCC-8 properties are as follows: externally connected (EC), disconnected (DC), covered by/tangential proper part (TPP), inside/non-tangential proper part (NTPP), equal (EQ), partial overlap (PO), covers/tangential proper part inverse (TPPi), and contains/non-tangential proper part inverse (NTPPi).

The spatial ontology consisted of three sub-ontologies: basic and two language ontologies. The basic ontology contained concepts and spatial properties. The two language ontologies contained English and Lithuanian toponyms. Words in language ontologies were matched with concepts in the basic ontology (e.g. *United States*, *US* and *USA* represent the same concept *USA*). All locations in language ontologies were represented by a *geo-info.owl* code and lexically represented by a *hasLexrep* relation.

A list of instances was created on the basis of the GeoNames database[6] (7 continents, 193 countries, 51 USA states, 6359 USA cities, 6955 Lithuanian place names, 1869 cities from top 10 cities of other countries). The GeoNames database contains information about continents, countries and cities and it contains information about spatial relations between these objects. RCC-8 relations were extracted from the GeoNames database.

[5] www.solim.eu
[6] www.geonames.org

To query the spatial ontology we used the function *GetSpatialRelations(A,B)* to get spatial knowledge about relations between A and B. This information can be inferred from the spatial ontology, whereas we cannot get false or unknown information, for example:

- *GetSpatialRelations(Georgia,Armenia)*= ”EC” only if there is enough information in the ontology to infer this relation;
- *GetSpatialRelations(Georgia,Latvia)*= ”DC” if this relation can be inferred;
- *GetSpatialRelations(Georgia, Latvia)*= ””, if there is not enough information in the ontology to infer the DC relation.

2.4 Implemented SMT System with Spatial Knowledge

For the implemented system with spatial knowledge we used the same training corpora as for the baseline system, as well as prepared two more corpora from the ontology – a translation dictionary (~0,02 mil. units) and spatial relation dictionary (~0,42 mil. units).

The developed baseline SMT system was a pure phrase-based SMT system which dealt only with surface forms of words. Its translation model contained simple probabilities like:

- *P(Georgia|Gruzija)* – a probability that *Georgia* is the English translation of the Lithuanian word *Gruzija*;
- *P(Georgia|Džordžija)* – a probability that *Georgia* is the English translation of the Lithuanian word *Džordžija*.

It also contained probabilities for all morphological variants of Lithuanian words and phrases. However, it was difficult to choose the correct Lithuanian translation of a given ambiguous English toponym since both probabilities were similar:

$$P(Georgia|Gruzija) \cong P(Georgia|Džordžija).$$

The factored phrase-based SMT (Koehn and Hoang, 2007) is an extension of the phrase-based approach. It contains an additional annotation at a lexical unit level. The lexical unit is no longer just a token, but a vector of factors that represent different levels of annotation. The training data (a parallel corpus) has to be annotated with additional factors. For instance, it is possible to add lemma or part-of-speech information on source and target sides.

The implemented SMT system was based on the Moses toolkit that features factored translation models allowing the integration of additional layers of data directly into the process of translation. Spatial knowledge was used during training and translation processes as additional semantic factors integrated with the source language data. All toponyms in the source text were analysed and tagged (annotated) with semantic factors (spatial knowledge) inferred from the spatial ontology with a reasoner. For example, a toponym *Georgia* is ambiguous: it can refer to the USA state or the Caucasian country. See the example sentences:

- There are Lithuanians living in Georgia, Florida and other states.
- Experts have failed to travel to Georgia at the Tbilisi airport.

In the first sentence *Georgia* refers to the USA state, while in the second one it refers to the Caucasian country. To resolve this type of ambiguity, spatial knowledge was used to determine spatial relations between corresponding toponyms within one sentence. For example, in the first sentence *Georgia* was annotated with *EC.Florida* since that information had been inferred from the spatial ontology (*Georgia* is externally connected to *Florida*). In the second sentence *Georgia* was annotated with *NTPPi.Tbilisi* (*Tbilisi* is a city in *Georgia*). We searched a sentence for toponyms and queried the spatial ontology for their relations. If there were more than two toponyms in a sentence we used just one (the first found, but not DC) annotation to each toponym. Compared with a simple unfactored translation model, that kind of factored translation model contained more useful information for toponym disambiguation since it might contain probabilities like:

- *P(Georgia/EC.Florida|Džordžija)* – a probability that *Georgia* is the English translation of a Lithuanian word *Džordžija* given that *Georgia* is externally connected to *Florida*;
- *P(Georgia/NTPPi.Tbilisi|Gruzija)* – a probability that *Georgia* is the English translation of Lithuanian word *Gruzija* given that *Georgia* encloses *Tbilisi*.

The translation model with probabilities about words and phrases with spatial knowledge helped to perform more accurate toponym disambiguation, because spatial context was included in the translation model. For example, if we have almost equal probabilities for *Georgia*, being a translation of both *Gruzija* and *Džordžija* in the translation model of the baseline system, probabilities with spatial knowledge are significantly different:

$$P(\text{Georgia/EC.Armenia}|\text{Gruzija}) \gg P(\text{Georgia/EC.Armenia}|\text{Džordžija})$$

$$P(\text{Georgia/EC.Florida}|\text{Džordžija}) \gg P(\text{Georgia/EC.Florida}|\text{Gruzija})$$

Thus, during the machine translation process semantic factors inferred from the spatial ontology provide additional information for the Moses decoder. As a result, it helps in choosing the appropriate translation equivalent. Therefore, SMT training data annotated with the proposed kind of spatial knowledge leads to a better machine translation quality.

It should also be mentioned that two SMT systems with spatial knowledge were trained. The first system (later referred as Spatial-8) was trained using corpora annotated with all eight RCC-8 spatial relations. The second system (later referred as Spatial-7) was trained using only seven RCC-8 relations since initial experiments, proved with the linguistic analysis, showed that using the *DC:disconnected* relation did not help in toponym disambiguation.

3 Evaluation and Limitations

A multifaceted strategy with three procedures was applied to the evaluation of the output quality of machine translation performed by the implemented system with spatial knowledge:

- automatic (black-box) evaluation;
- human evaluation;
- linguistic analysis.

3.1 Automatic Evaluation

For the automatic evaluation the two most popular and widely used metrics BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) were used. Automatic metrics are cost-effective and do not require much human intervention. They allow comparisons of two and more systems, as well as different versions of one system in the process of its implementation and improvement as many times as necessary.

A balanced test set of 500 English sentences was developed for the automatic evaluation purposes. Sentences were manually collected from

the web and translated into Lithuanian by a professional translator (a reference set to be compared with). The breakdown of topics in the corpus is presented in Table 2.

Domain	Percentage
General information about the EU	12%
Specification and manuals	12%
Popular scientific and educational	12%
Official and legal documents	12%
News and magazine articles	24%
Information technology	18%
Letters	5%
Fiction	5%

Table 2. Testing set.

The procedure of the automatic evaluation consists of several sub-processes and the main idea, in general, is in the comparison of machine translation and reference sets. The higher the automatic scores are, the better the machine translation output quality is. BLEU and NIST scores for the baseline system were 27,35 and 5,90 correspondingly. BLEU and NIST scores for the implemented system with spatial knowledge were 27,97 (BLEU) and 5,97 (NIST) for the system "Spatial-8" and 27,47 (BLEU) and 5,91 (NIST) for the system "Spatial-7" (see Table 3).

System	BLEU	NIST
Baseline	27,35	5,90
Spatial-8	27,97	5,97
Spatial-7	27,47	5,91

Table 3. Results of the automatic evaluation.

As a result, a slight improvement in the output quality of machine translation with spatial knowledge can be observed. In general, this improvement is not high and is not sufficient for the objective and an integrated evaluation procedure. Results of the automatic evaluation can be explained so that general-purpose development and evaluation corpora used for the evaluation did not contain many ambiguous geographical names. Therefore, the evaluation with the task-specific evaluation corpus was performed during the human evaluation. Nevertheless, automatic scores were set as a threshold for further experiments.

3.2 Human Evaluation

A test set of 464 English sentences containing ambiguous toponyms was developed for human evaluation purposes. A ranking of translated sentences relative to each other was used for the manual evaluation of systems. This was the official determinant of translation quality used in the 2009 Workshop on Statistical Machine Translation shared tasks (Callison-Burch et al., 2009).

A web-based human evaluation environment (Skadiņš et al., 2010) was used where source sentences and translation outputs of the two SMT systems could be uploaded as simple txt files. Once the evaluation of the two systems was set up, a link to the evaluation survey was sent to evaluators. Evaluators were evaluating the systems sentence by sentence. Evaluators saw the source sentence and the translation output of the two SMT systems – baseline and the one implemented with spatial knowledge. The frequency of preferring each system based on evaluators' answers and a comparison of the sentences was calculated. About 20 evaluators participated, each comparing translations of 50 sentences.

The manual comparison of the two systems (Baseline vs. Spatial-8)[7] has shown that the implemented SMT system with spatial knowledge is slightly better than the baseline system: in 50,66% of cases evaluators judged its output to be better than the output of the baseline system. Results of the human evaluation do not allow us to say with certainty either the spatial SMT system is significantly better or it is disambiguating toponyms better, since the difference is not convincing and evaluators have been comparing sentences using subjective criteria and not paying a special attention to the translation of toponyms.

3.3 Linguistic Evaluation of Toponym Disambiguation

A detailed linguistic analysis of toponym disambiguation during the machine translation process was performed. The same corpus as for the human evaluation was used and the accuracy of the toponym translation was evaluated. The accuracy of the baseline system was 84,09%. The accuracy of the Spatial-8 system was 83,87%. Since results for the baseline system were better, it was decided to analyse the impact of each spatial relation to toponym disambiguation. It was discovered that the accuracy could be increased to 88.00% if the DC:*disconnected* relation was ignored (system Spatial-7).

[7] The human evaluation of the system Spatial-7 is in progress at the moment and will be presented in the final version of the paper.

4 Conclusions and Future works

In the paper we have presented how toponyms can be disambiguated in the process of statistical machine translation using spatial knowledge by the example of the English-Lithuanian system. We have overviewed the system design including the description of its functionality, baseline and implementation with spatial knowledge, as well as focused on the system multifaceted evaluation and its results.

We can see that the quality of machine translation can be improved by using the semantic information from the spatial ontology. Nevertheless improvement is not big and further more detailed evaluation would be necessary to assess whether this improvement is statistically significant.

It was noticed during linguistic evaluation that some RCC-8 properties seem to be much more useful than others (e.g. *EC:externally connected* and *EQ:equal*). But a detailed evaluation of the impact of each relation has not been done yet. The EQ property can be used for machine translation of toponyms which are synonyms, for example, a full name and an abbreviation – *the United States of America* and *USA*. The same property can be used for the so-called exonyms (names of places used by other groups, not locals) as *Praha* for its inhabitants and *Prague* for the English (for other examples, see Leidner (2007)).

It should be also noted, that the best version of the implemented system with the spatial ontology is not dealing with *DC*:*disconnected* relations, e.g. *Georgia* is disconnected from *California* or *Hawaii*. In this case, other types of information in the spatial ontology may be used in further experiments, e.g. the ontology class *State* and its instances.

Moreover, the spatial ontology was not used for disambiguation of common nouns since they were not represented in the ontology. However, a morpho-syntactic type of toponym ambiguity, when a word itself can be a toponym or a common noun in a language) and its resolution can be performed with the help of the spatial ontology, for example:

- *Hook* refers to the populated place in the UK and *hook* is a common noun (English);
- *Liepa* refers to the populated place in Latvia and *liepa* (lime-tree) is a common noun (Latvian);
- *Batą* refers to the populated place in Lithuania and *batą* (shoe) is a common noun (Lithuanian).

The proposed approach to toponym disambiguation is not limited to:

- machine translation *per se* and can be regarded as generic, i.e. it can be also applied to other fields of natural language processing, e.g. information retrieval;
- use of spatial knowledge only: other types of implicit or inferred knowledge can be used in a similar way.

References

Brown P., Della Pietra S., Della Pietra V., Mercer R. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), pp. 263–311.

Callison-Burch C., Koehn P., Monz C., Schroeder J. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 1-28, Athens, Greece.

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

Koehn P., Och F. J., Marcu D. 2003. Statistical phrase based translation. 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).

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

Li Z., Callison-Burch C., Dyer C., Ganitkevitch J., Khudanpur S., Schwartz L., Thornton W., Weese J., Zaidan O. 2009. Joshua: An Open Source Toolkit for Parsing-based Machine Translation. *Proceedings of the Workshop on Statistical Machine Translation (WMT09).*

Marcu D., Wang W., Echihabi A., Knight K.. 2006. SPMT: statistical machine translation with syntactified target language phrases. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, July 22-23. ACL Workshops. Association for Computational Linguistics, pp. 44-52.

Quirk C., Menezes A., Cherry C. 2005. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. *Proceedings of ACL 2005*.

Papineni K., Roukos S., Ward T. et al. 2002. BLEU: a Method for Automatic Evaluation of Machine

Translation. *ACL'02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, Pensilvania. Morristown, NJ: Association for Computational Linguistics, pp. 311-318.

Doddington G. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *HLT 2002: Human Language Technology Conference: Proceedings of the Second International Conference on Human Language Technology Research*, San Diego, California. San Francisco: Morgan Kaufmann Publishers, pp. 138-145.

Tiedemann J. and Nygaard L. 2004. The OPUS corpus – parallel & free. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 26-28.

Tiedemann J. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *Recent Advances in Natural Language Processing*, vol. V, John Benjamins, Amsterdam/Philadelphia, pp. 237-248.

Randell D. A., Cui Z., Cohn A. G. 1992. A spatial logic based on regions and connection. *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, Morgan Kaufmann, San Mateo, pp. 165–176.

Skadiņš R., Goba K. and Šics V. 2010. Improving SMT for Baltic Languages with Factored Models. *Proceedings of the Fourth International Conference Baltic HLT 2010*, Riga, Latvia, pp. 125-132.

Och F. J. and Ney H. 2003 A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, (29)1, pp. 19-51, 2003.

Koehn P., Federico M., Cowan B., Zens R., Duer C., Bojar O., Constantin A., Herbst E. 2007 Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, pp 177-180.

Madnani Nitin, Resnik Philip, Dorr Bonnie, Schwartz Richard. 2008. Applying Automatically Generated Semantic Knowledge: A Case Study in Machine Translation. *Proceedings of the Symposium on Semantic Knowledge Discovery, Organization and Use.*

Leidner Jochen L.. 2007. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* PhD thesis. Institute for Communicating and Collaborative Systems School of Informatics, University of Edinburgh.

Gornostay T. and Skadiņa I. 2009. English-Latvian Toponym Processing: Translation Strategies and Linguistic Patterns. *EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, May 14-15, Universitat Politècnica de Catalunya, Barcelona, Spain, pp. 81-87.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955