English-Latvian Toponym Processing:

Translation Strategies and Linguistic Patterns

Tatiana Gornostay
Tilde, Latvia
tatjana.gornostaja@tilde.lv

Inguna Skadiņa
Tilde, Latvia
inguna.skadina@tilde.lv

Abstract

The paper presents a study of a challenging task in machine translation and cross-language information retrieval – translation of toponyms. Due to their linguistic and extra-linguistic nature, toponyms deserve a special treatment. The overall translation process includes two stages of processing: dictionary-based and out-of-vocabulary toponym translation. The latter is divided into three steps: source string normalisation, translation, and target string normalisation. The translation process implies an application of translation strategies and linguistic toponym translation patterns. Possible translation strategies, including transliteration and translation *per se* along with combined strategies, and linguistic toponym translation patterns, including multi-word patterns as well, were investigated and implemented for English-Latvian machine translation. 10,000 The UK-related toponyms from Geonames were selected for a development set. The evaluation of output quality on basis of a test set has showed 67% accuracy in out-of-vocabulary translation: 58% on a set containing one-word toponymic units and 81% on a multi-word test set.

1 Introduction

The paper presents a study of a challenging task in machine translation (MT) and cross-language information retrieval (CLIR) – translation of toponyms. Due to their linguistic and extra-linguistic nature, toponyms deserve a special treatment.

In general toponyms are studied by toponymy and represent names of places comprising the following types:

- *hydronyms* (names of bodies of water: bays, streams, lakes, lagoons, oceans, ponds, seas, etc., e.g. *Thames* as a river);
- *oronyms* (names of mountains, cliffs, craters, rocks, points, etc., e.g. *Bexhill* as a mountain);
- *geonyms* (general names for streets, squares, lines, avenues, paths, alleys, roads, embankments, etc.);
- *oeconyms* (names of populated places: an administrative division, country, city, town, house or other building).

The first part of the paper overviews the concept and nature of toponyms along with existing toponym translation strategies (TS). The second part of it focuses on the developed and implemented English-Latvian toponym MT approach, including a description of TSs and linguistic toponym translation patterns (LTTP).

2 Concept and Nature of Toponyms

After Geoffrey Leech (1981) we can accept a special status of toponyms as proper names without a conceptual meaning as we cannot perform any componential analysis for them. However, we cannot but admit the fact that many toponyms are at least meaningful etymologically, e.g, *Cambridge* – bridge over the river *Cam* (Leidner, 2007), and, as Leidner pointed out, this etymology might or might not be apparent to a speaker. This feature makes toponyms difficult for processing.

Proceedings of the 13th Annual Conference of the EAMT, pages 81–87,
Barcelona, May 2009

Besides, toponyms are not unambiguous. Leidner (2007) describes three types of the toponymical ambiguity:

- *morpho-syntactic* ambiguity: a word itself may be a toponym or may be a common noun in a language, e.g. *Hook* as the populated place in the UK versus *hook* as a common noun;
- *referential ambiguity*: a toponym may refer to more than one place of the same type, e.g. *Riga* as the populated place and the capital of Latvia and *Riga* as the populated place in the USA, state Michigan;
- *feature type ambiguity*: a toponym may refer to more than one place of different type, e.g. *Tanfield* as the populated place and the castle in the UK, *Gauja* as the populated place and the river in Latvia.

Moreover, there is another type of the toponymical ambiguity to be addressed, that is the so-called *eponymical ambiguity* when names of places are named after people or deities, e.g. *Vancouver* after George Vancouver. In addition, the same place is often known by different names – *endonyms* (names of places used by inhabitants, self-assigned names) and *exonyms* (names of places used by other groups, not locals) as in the Leidner's (2007) example with *Praha* for its inhabitants and *Prague* for English.

Furthermore, metonymy also contributes to the issue. This linguistic phenomenon was studied from the toponymical point of view by Markert and Nissim (2002). The authors stated that the metonymic use of toponyms is regular and productive, can reach up to 17% of all of toponyms as it was proved by the example of the English language, and the most frequent and conventional case of the toponymical metonymy is as in the "*government of ...*" pattern, e.g. "*Latvia announced ...*" means "*the government of Latvia announced ...*".

Finally, toponyms are changed frequently since they themselves and the places they refer to are not constant. Therefore, when dealing with toponyms it is also very important to take into consideration historical and cultural facts.

The abovementioned linguistic and extra-linguistic features make toponym processing difficult, e.g. resolution, retrieval, and especially translation.

3 Toponym Translation Strategies and Approaches

Toponyms can be referred to *named entities* (NE) which comprise all types of proper names, including toponyms themselves, anthroponyms, and temporal expressions. To translate an NE one should choose a TS which depends on the type of the NE unit (Babych and Hartley, 2003), i.e. to translate a toponym we should know its type that assigns a TS to be applied to this toponym. Common TSs for toponyms, as a particular type of NEs, are the following (Babych and Hartley, 2004):

- *transference* strategy, i.e. do-not-translate;
- *transliteration* strategy, i.e. phonetic or spelling rendering;
- *translation* strategy *per se*, i.e. do-translate;
- *combined* strategy, i.e. applying more than one from the abovementioned three strategies.

The transference strategy with the do-not-translate list is often used for translation of toponyms which do not need any rendering at all and are often left not translated, e.g. organization names (Babych and Hartley, 2003).

The topic of transliteration has been studied for several languages, mostly for non-Latin spelling, and many techniques have been proposed. The most common transliteration techniques are phoneme-based and grapheme-based (Zhang et al., 2004). The phoneme-based approach (Knight and Graehl, 1998; Meng et al., 2001; Oh and Choi, 2002; Lee and Chang, 2003) implies a conversion of a source language word into a target language word via its phonemic representation, i.e. grapheme-phoneme-grapheme conversion. The grapheme-based technique converts a source language word into a target language word without any phonemic representation (grapheme-grapheme conversion) (Stalls and Knight, 1998; Li et al., 2004).

Most of toponym translation approaches are data-driven (see, e.g. Meng et al., 2001; Al-Onaizan and Knight, 2002; Sproat et al., 2006; Alegria et al., 2006; Wentland et al., 2008) since they deal with widely used languages which have enough linguistic resources for development. Taking into account an under-resourced status of the Latvian language with few available corpus resources, especially parallel bilingual corpora, a rule-based approach was proposed for English-Latvian toponym translation.

4 Implementation of English-Latvian Toponym Translation

Strategies and techniques for English-Latvian toponym MT have not been studied previously, and the existing literature describes general principles of rendering of the English proper names, mostly anthroponyms, into Latvian.

We studied three main issues of English-Latvian toponym MT:

- orthographic, phonetic and grammatical differences between the two languages;
- possible toponym translation strategies for this translation direction;
- possible linguistic toponym translation patterns for this translation direction.

Although English and Latvian are the Indo-European languages and share some grammatical features, they have a lot of differences since English belongs to the Germanic language group while Latvian belongs to the group of the Baltic languages; English is an analytical language in contrast to the synthetic Latvian language with a rich set of inflections and some specific orthographic features such as diacritics. The lack of the orthographic and phonetic convergence in English (26 letters to 44 phonemes), historical changes and traditions in spelling, origin language of a toponym, and ambiguity, as well as the lack of the Latvian linguistic resources for the study were the main difficulties we faced. We also studied the peculiarities of Latvian toponymic units to ensure they correspond to the Latvian grammar and orthography rules, e.g.:

- Latvian names are inflected;
- Latvian names cannot be spelled with double consonants, except *ll*, *mm* or *nn* under certain conditions;
- Latvian multi-word units can be translated in several ways, however, a compound is preferable if it allows to reconstruct a source toponymic unit (Ahero, 2006).

4.1 Source String Normalisation

Translation of a toponymic unit is divided into three steps: source string normalisation, translation, i.e. application of TS and LTTP, and target string normalisation according to the Latvian grammar and orthography rules.

Source string normalisation includes the following sub-processes:

- all tabs and double space characters, including the beginning of a string, are normalized to single space characters;
- the so-called "zero-fertility words" (Al-Onaizan and Knight, 2002) of English are normalized to zero-translations in Latvian, e.g. definite article *the* is omitted;
- hyphenated words are normalized to non-hyphenated ones;
- some abbreviations are expanded to full words, e.g. *St.* to *Saint*;
- signs, if possible, are changed to words, e.g. *&* to *and*;
- punctuation marks are normalized to zero translations.

4.2 Translation: English-Latvian Toponym Translation Strategies

Transference strategy is applied to unprocessed toponymic units which are not described by any of LTTPs.

Transliteration strategy is language dependent (Karimi et al., 2007) and for the English-Latvian language pair transliteration is a non-trivial task due to differences in grammar, orthography and sound systems of both languages. Moreover, there are a lot of exceptions (see Castañeda-Hernández, 2004 about general toponym translation problem). English-Latvian transliteration strategy is based on the grapheme-to-grapheme approach, which implies direct mapping of the English letter sequences into the Latvian ones, formalized in a set of transliteration rules. All foreign names (those of non-English origin) are rendered according to the English pronunciation standards. The main principle is the possibility to reconstruct a source toponymic unit (Ahero, 2006).

The set of English-Latvian transliteration rules consists of about 110 transliteration patterns describing English-Latvian grapheme-to-grapheme correspondences. The result of transliteration may vary, as there can be several ways of rendering the English letter combinations into the Latvian ones. Several cases of variety are described by transliteration patterns, e.g. *-c-* stands for *-k-* before consonants (except *-h-*), and *-a-*, *-o-*, *-u-*, for *-s-* before *-i-*, *-e-*, *-y-*, and for *-č-* in the combination with *-h-*.

Translation strategy *per se* is also applied to English-Latvian toponym translation. In some cases toponyms are not transferred or transliterated, but translated into Latvian, e.g. multi-word units *East Anglian Heights, North West Highlands* are translated into Latvian as *Austrumanglijas augstiene*, *Ziemeļskotijas kalnāji* correspondingly, while one-word units are transliterated, as a rule. Though, transliteration strategy can be also applied to multi-word units in parallel with translation one which is usually infrequent and conventional.

Toponym TSs are closely related with LTTPs and are language dependent. Therefore, *combined* strategy is also used when treating different types of toponyms.

4.3 Translation: Linguistic Toponym Translation Patterns

When translating a toponymic unit, dictionary-based translation is applied first. Most of popular toponyms, such as names of countries and capitals, seas and oceans, are translated using an English-Latvian dictionary, e.g. *Lisbon – Lisabona*, *Brussels – Brisele*, *Cologne – Ķelne*, *Antwerp – Antverpene*, *Great Britain – Lielbritānija*, *Atlantic Ocean – Atlantijas okeāns*. If a toponym is an out-of-vocabulary (OOV) word then one of the LTTPs is applied.

To determine possible LTTP we studied a list of 10,000 toponyms from Geonames (all toponyms were UK-related) and analyzed 59 toponym types.

Generally, LTTPs are the ways source toponymic units are rendered into target toponymic units. LTTPs can be of two types: in-word patterns and multi-word patterns. The in-word LTTP is a word transformation model, based on English-Latvian transliteration rules, including the most frequent prefixes, suffixes, and letter combinations. There are about 300 in-word LTTPs described, for example: *new-* to *ņū-*, *deep-* to *dīp-*, *mc-* to *mak-*, *-worth* to *–vērt, -islet* to *–ailet*, etc.

Multi-word LTTPs involve three TSs. Translation strategy S_1 is based on transliteration rules. Translation strategy S_2 performs the combination of the first TS and the insertion of a nomenclature word, e.g. *Bebington* (as a railroad station) – *Bebingtonas* ***stacija***. If a nomenclature word is included in a source toponymic unit, as it is in the pattern S_3, it is either translated (*Newton* ***Point*** - *Ņūtona* ***zemesrags***, *Gog Magog* ***Hills*** - *Gogmagogu* ***kalni***) or transliterated (*Green* ***Isle*** – *Grīn**aila***, *North East* ***Coast*** – *Nortīst**kosta***) in a target language. We described 40 nomenclature words that are translated under certain conditions. Auxiliary words, such as prepositions, are also either translated or transliterated, e.g. *Horse* ***of*** *Copinsay – Hors**of**kopinsejs* (transliteration), *Milford* ***upon*** *Sea - Milforda* ***pie*** *jūras* (translation).

Examples of LTTP are presented in Table 1. X_n is a toponymic unit in a source language, S_n is a translation strategy applied, Y_n is a toponymic unit in a target language, and $P_n\{X_n, S_n, Y_n\}$ is a corresponding LTTP.

4.4 Target String Normalisation

Target string normalisation modifies a toponymic unit according to the rules of the Latvian grammar and orthography, e.g. all populated places are feminine gender (see P1): *Newcastle* → *Ņūkāsl**a*** which is indicated by the ending *–a* (feminine, singular nominative).

English Toponym X_n	**Translation Pattern P_n**	**Translation Strategy S_n**	**Latvian Toponym Y_n**
$P_1\{X_1, S_1, Y_1\}$			
X1: N *Knocklayd*	P1: N → N	S1: transliteration	Y1: N masculine singular *Nokleids*
$P_2=\{X_1, S_1, Y_2\}$			
X1: N *Newcastle*	P2: N → N	S1: transliteration	Y2: N feminine singular *Ņūkāsla*
$P_3=\{X_1, S_2, Y_3\}$			
X1: N *Bebington*	P3: N → N + N	S2: transliteration + nomenclature word	Y3: N feminine singular genitive + N *Bebingtonas stacija*
$P_4=\{X_2, S_1, Y_2\}$			
X2: N’s + N *Bishop's Stortford*	P4: N’s + N → N	S1: transliteration	Y2: N feminine singular *Bišopsstortforda*

$P_5=\{X_3, S_1, Y_2\}$			
X3: N + N's + N *St. Bishop's Town*	P5: N + N's + N → N	S1: transliteration	Y2: N feminine singular *Sentbišopsatauna*
$P_6=\{X_4, S_1, Y_2\}$			
X4: N + N *Bishop Auckland* *North Ronaldsay*	P6: N + N → N	S1: transliteration	Y2: N feminine singular *Bošopoklenda* *Nortronaldseja*
$P_7=\{X_5, S_1, Y_2\}$			
X5: A + N *South Ribble, Green Isle*	P7: A + N → N	S1: transliteration	Y2: N feminine singular *Sautribla* *Grīnaila*
$P_8=\{X_6, S_3, Y_4\}$			
X6: N + P + N *Milford upon Sea* *Stratford upon Avon*	P8: N + P + N → N + P + N	S3: transliteration + translation	Y4: N feminine singular genitive + P + N *Milforda pie jūras,* *Stradforda pie Avona*
$P_9=\{X_6, S_1, Y_5\}$			
X6: N + P + *Longville in the Dale*	P9: N + P + N → N + N	S1: transliteration	Y5: N feminine singular genitive + N feminine singular locative *Longvila Deilā*
$P_{10}=\{X_7, S_1, Y_2\}$			
X7: A + A + N *North East Coast*	P10: A + A + N → N	S1: transliteration	Y2: N feminine singular *Nortīstkosta*
$P_{11}=\{X_8, S_2, Y_3\}$			
X8: N + C + N *Sandal & Agbrigg*	P11: N + C + N → N + N	S2: transliteration + nomenclature word	Y3: N feminine singular genitive + N *Sendalendagbrigas stacija*
$P_{12}=\{X_4, S_3, Y_6\}$			
X4: N + N *Newton Point*	P12: N + N → N + N	S3: transliteration + translation	Y6: N masculine singular genitive + N *Ņūtona zemesrags*
$P_{13}=\{X_6, S_1, Y_1\}$			
X6: N + P + N *Horse of Copinsay*	P:13 N + P + N → N	S1: transliteration	Y1: N masculine singular *Horsofkopinsejs*
$P_{14}=\{X_7, S_3, Y_7\}$			
X7: N + N + N *Gog Magog Hills*	P14: N + N + N → N + N	S3: transliteration + translation	Y7: N masculine plural genitive +N *Gogmagogu kalni*

"Table 1. Examples of English-Latvian Linguistic Toponym Translation Patterns."

5 Evaluation and Limitations

The current MT evaluation theory and practice lacks in evaluation methods for toponym translation task. One of the reasons could be that it is not clear what the correct toponym translation is, since results may vary and more than one target toponymic unit is acceptable. As a result, scores calculated with a single target variant will underestimate translation accuracy. Moreover, human translations are often inaccurate as well.

Existing English-Latvian MT systems[2] do not implement any OOV algorithms to translate toponymic units. Thus, we had no possibility to

[2] English-Latvian Pragma Expert: www.acl.lv, English-Latvian Google: http://translate.google.com, English-Latvian Tilde http://www.tilde.lv/English/portal/go/tilde/3777/en-US/DesktopDefault.aspx (November, 2008)

compare our algorithm with other MT performance.

For evaluation purposes we compared translation results of our translation module with reference (human) translations from two bilingual dictionaries. 330 English toponymic units of different types with Latvian translation equivalents were manually extracted from dictionaries (180 one-word units and 150 multi-word units) and processed with our OOV toponym translation algorithm. To evaluate translation results we set the following scores:

- if the translation result coincides with the corresponding linguistic toponym translation pattern then the translation is *accurate* and the score is 1;
- if the translation result deviates from the corresponding linguistic toponym translation pattern then the translation is *inaccurate*, and the score is 0,5 for one error and 0 for more errors.

We accept variants as they were also described by LTTPs (in transliteration rules). As a result, the accuracy of translation is 67% on the whole test set, 58% on the set containing one-word toponymic units, and 81% on the multi-word test set.

6 Conclusions and Future Work

We have described the pattern-based toponym translation approach developed for the English-Latvian language pair. We studied the concept and nature of toponyms and several linguistic and extra-linguistic issues, such as ambiguity, cultural and historical changes and others. We also studied different types of toponyms in the context of the overall task of toponym MT.

In the present paper we have overviewed two stages of toponym translation processing: dictionary-based and OOV toponym translation. The latter is divided into three steps: source string normalisation, translation and target string normalisation. The focus of the paper is on detailed description of OOV toponym processing: possible translation strategies and linguistic toponym translation patterns with examples and evaluation results.

We can conclude that for the implemented rule-based approach there is much room for possible improvements, and evaluation results prove this statement. The main reason, why toponym processing is such a challenging task for MT, is the necessity of knowledge of toponym rendering rules, variety of languages as well as a considerable amount of history and culture (Castañeda-Hernández, 2004). It is impossible to formalize this process completely and it is obvious that there can be mistakes in automated translation of toponymic units.

Corpus-based approach has not been applied in this research due to the lack of monolingual and bilingual linguistic resources. However, the issue of compiling a corpus of toponym-referenced texts for the Latvian language is being studied. We also plan to study the issue of multilingual cross-language toponym MT and application MT strategies to other languages (especially Cyrillic or other non-Latin scripts).

Acknoledgement

The research reported was partly funded by the TRIPOD project (TRI-Partite multimedia Object Description)[3] supported by the European Commission under the contract No. 045335.

We would like to thank Lars Ahrenberg, Lars Borin and Raivis Skadiņš for discussions and comments.

References

Antonija Ahero. 2006. *English Proper Name Rendering into the Latvian Language* (Angļu Īpašvārdu Atveide Latviešu Valodā). Zinātne, Rīga.

Iñaki Alegria, Nerea Ezeiza, Izaskun Fernandez. 2006. Named entities translation based on comparable corpora. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Workshop on Multiword expressions in a Multilingual Context*, Italy. Pp.1-8.

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, USA. Pp.400-408.

Bogdan Babych and Anthony Hartley. 2003. Improving Machine Translation Quality with Automatic Named Entity Recognition. *Proceedings of the 7th European Association for Machine Translation Workshop Improving machine translation through other language Technology Tools*, Hungary. Pp.1-8.

Bogdan Babych and Anthony Hartley. 2004. Selecting Translation Strategies in MT using Automatic Named Entity Recognition. *Proceedings of the 9th European Association for Machine Translation*

[3] http://tripod.shef.ac.uk/

Workshop Broadening horizons of machine translation and its applications, Malta. Pp.18-25.

Gilberto Castañeda-Hernández. 2004. Navigating through Treacherous Waters: The Translation of Geographical Names. *Translation Journal*, 8(2): [electronic resource]: http://accurapid.com/journal/28names.htm#1

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2007. Collapsed consonant and vowel models: new approaches for English-Persian transliteration and back-transliteration. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Czech Republic. Pp.648-655.

Kevin Knight and Jonathan Graehl. 1998. Machine Transliteration. *Computational Linguistics*, 24(4):599-612.

Chun-Jen Lee and Jason S. Chang. 2003. Acquisition of English-Chinese Transliteration Word Pairs from Parallel-Aligned Texts using a Statistical Machine Translation Model. *Proceedings of Human Language Technologies – The North American Chapter of the Association for Computational Linguistics Workshop: Building and Using parallel Texts Data Driven Machine Translation and Beyond*, Canada. Pp.96-103.

Geoffrey Leech. 1981. *Semantics. The Study of Meaning.* 2nd edition. Penguin, London, England, UK.

Jochen L. Leidner. 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.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. *Proceedings of the 42nd Annual Meeting on association for Computational Linguistics*. Spain. Pp.159–166.

Katja Markert and Malvina Nissim. 2002. Towards a corpus annotated for metonymies: the case of location names. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, France. Pp.1385-1392.

Helen M. Meng, Wai-Kit Lo, Berlin Chen, and Karen Tang. 2001. Generate Phonetic Cognates to Handle Named Entities in English-Chinese cross-language spoken document retrieval. *Proceedings of Institute of Electrical and Electronics Engineers Automatic Speech Recognition and Understanding Workshop*, Italy.

Jong-Hoon Oh and Key-Sun Choi. 2002. An English-Korean Transliteration Model Using Pronunciation and Contextual Rules. *Proceedings of the 19th International Conference on Computational Linguistics*, Taiwan, 1:1-7.

Richard Sproat, Tao Tao, and Cheng-Xiang Zhai. 2006. Named entity transliteration with comparable corpora. *Proceedings of the 44th Annual meeting of the Association for Computational Linguistics*, Australia. Pp.73-80.

Bonnie Glover Stalls and Kevin Knight. 1998. Translating Names and Technical Terms in Arabic Text. *Proceedings of the Coling / Association for Computational Linguistics Workshop on Computational Approaches to Semitic Languages*, Canada. Pp.365-266.

Wolodja Wentland, Johannes Knopp, Carina Silberer, and Matthias Hartung. 2008. Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration. *Proceedings of the 6th Language Resources and Evaluation Conference*, Morocco.

Min Zhang, Haizhou Li, and Jian Su. 2004. Direct Orthographical Mapping for Machine Transliteration. *Proceedings of the 20th International Conference on Computational Linguistics*, Switzerland.