Ontology-based terminology management for transitive translations focusing on NEs

Fumiko Kano Glückstad
Department of International Language Studies and Computational Linguistics
Copenhagen Business School
Dalgas Have 15, DK-2000 Frederiksberg, Denmark
fkq.isv@cbs.dk

I demonstrate that there are two types of transitive translations of Named Entities (NEs), both of which should be handled in the process of Cross Lingual Information Retrieval (CLIR). An official transitive translation is defined as a translation made through an official English translation often provided from an authorized local entity. A lexical translation is a direct lexical translation from a source language to a target language. However, a lexical translation also requires a transitive translation when a language pair is a rare combination having inadequate language resources. Hence I define it as a lexical transitive translation. I assess the inconsistency level of the official- and lexical transitive translations of NEs and propose an ontology-based CLIR solution referred to as *triangulated terminology management*.

Multilingual information retrieval, CLIR, transitive translation, pivot translation, named entity disambiguation

1. BACKGROUND

My research issue has been raised by the question: Is it possible to identify local first-hand information produced in non-English speaking countries from Japanese queries translated from their official English information sources? Specifically, the issue is rooted in a plurality of inconsistencies found between Japanese translations made through the direct lexical translation from Danish to Japanese and Japanese translations made through the transitive translation using official English translations as source. A typical example of such a translation problem is illustrated where the formal English name of the Danish authority “Økonomistrelsen” is “The Danish Agency for Governmental Management.” The Danish originated name, “Økonomistrelsen”, will most likely be translated into a completely different Japanese expression through lexical English translations, “Economy Agency (*keizai-tyou*)” using available language resources such as Danish-English and English-Japanese dictionaries. Eventually, it becomes increasingly difficult for Japanese readers to identify the original Danish NE in the process of CLIR due to inconsistent Japanese translations. Hence my research addresses the following three issues: 1) evaluation of the inconsistency level between the direct Japanese translation of Danish NEs and of their officially translated English; 2) developing an ontology-based solution to identify an original entity from inconsistent translations based on a triangulated terminology management approach; and 3) eventually and hopefully, to identify a base-line CLIR system that can integrate the triangulated terminology management approach. This paper is only addressing the initial phases raised by issue 1.

2. THE OFFICIAL TRANSITIVE TRANSLATION AND LEXICAL TRANSITIVE TRANSLATION

In CLIR, one of the basic methods in query translation is called dictionary-based translation. The problem with this method is that there are inherently insufficient language resources available for most language pairs that are part of rare combinations. Hence, it is required to employ a transitive translation technique using a so-called pivot language. Gollins and Sanderson [1] pointed out that, since a transitive translation in CLIR is based on a simple word-by-word translation approach, it increases the likelihood of translation errors, caused mainly by incorrect identification of the sense of ambiguous words. Ballesteros [2] examined the impact of transitive translations and discovered that using simple word-by-word transitive translations from Spanish to French via English degraded performance by 91% when compared to a direct bilingual translation from Spanish to French. Gollins and Sanderson [1] introduced an approach to reducing errors by combining translations from two different transitive routes, a process known as *lexical triangulation*. Their results showed that the lexical triangulation approach to the transitive translation eliminated the differences in retrieval between transitive translated queries and equivalent direct translated queries.

However, considering the aforementioned specific example of the Danish NE, “Økonomistrelsen”, there are two types of transitive translations and the solution proposed by Gollins and Sanderson [1] only addresses issues arising from the lexical translation from Danish to Japanese. It means that it is necessary to clearly distinguish the transitive translation using an official English translation as inter-lingua from the transitive translation based on a

lexical translation. Hence in my research, I define the transitive translation using an official English translation as an *official transitive translation* and the transitive translation based on a lexical translation as a *lexical transitive translation*. In this work, I report the preliminary survey of measuring frequency and semantic similarity of the official- and the lexical transitive translations of Danish NEs.

In order to identify inconsistencies between the *official-* and *lexical transitive translation* of original Danish NEs, I compared differences between official English translations and lexical English translations of names of Danish governmental organizations (i.e. ministries and institutions under the ministries), most of which provide official English names of their organizations. For performing a *lexical translation* of original Danish NEs into English, I used one of the most popular Danish-English dictionary series entitled “Gyldendals Røde Ordbøger”. Regarding the *lexical translation*, I defined the following rules: 1) NEs consisting of several words should be translated word-by-word; 2) If the dictionaries propose an English translation equal to the corresponding official English translation, the official English expression should be applied. Accordingly, I translated all of 70 selected Danish NEs into English and extracted 26 English *lexical translations* that were not identical to their respective *official translations*. Since these English translations of NEs are Multi-Word Expressions, I further decomposed them into each lexical unit (word) and enlisted the inconsistent word pairs that were scope for further analysis. For comparing the semantic similarity of these word pairs, I used the basic Path Length measure provided on the web interface of the WordNet::Similarity [3]. The results showed the semantic distance in most of the word pairs produced via official- and lexical English translations. For example the official English expression of “Ministeriet for Videnskab, Teknologi og Udvikling (Ministry for Science, Technology and Development)” is “Ministry of Science, Technology and Innovation”. In the same way, “Forsknings- og Innovationsstyrelsen (Research and Innovation Agency)” is “Agency for Science, Technology and Innovation”. The semantic distances of word pairs “innovation vs. development” and “research vs. science” are respectively shown in FIGURE1.

FIGURE 1: Examples of Semantic similarity

4. OUTLOOK

My study shows that the similarity measures based on Path Length indicate the degree of inconsistency level between English translations made through a so-called *official translation* and a so-called *lexical translation*. The next noteworthy question is how a Japanese translation of these pairs of English translations will turn out. My initial assumption is that these Japanese translations will create expressions with an even deeper level of inconsistency (i.e. FIGURE 2). It means that it will be increasingly difficult to identify the original Danish NEs from various Japanese translations. If there were universal rules defining “a name should always be translated based on the lexical meaning of its original language”, these inconsistencies would potentially be tremendously reduced. However, the reality is unfortunately far from that. Usually, the decision of names and their translations involves a plurality of issues, such as political (domestically, internationally), cultural, social and so on. It means that problems originating from both *official-* and *lexical transitive translations* should be carefully dealt with in terms of a so-called Named Entity Disambiguation.

As a solution, I propose an ontology-based triangulated terminology management approach. The approach is based on the idea that a country specific NE has a unique ontological structure, since a named entity is per definition unambiguously defined on a global scale. For example, the Danish governmental organizations are existing according to a Danish governmental structure that is uniquely defined in this country. It means that the ontological structure is unique even though each named entity is expressed in different languages. Therefore, an ontology-based terminology database consists of three layers: a) each NE expressed in a source language, b) its official expression in an inter-lingual language (usually in English), and c) all possible expressions in a target language (FIGURE 3). Each entity in an ontological hierarchy should contain metadata specifying country, timeframe, structural relation etc. These three layers should have a triangulated relationship as shown in FIGURE 4. The key issue is that the name of an entity expressed in a source language and an official expression in an inter-lingual language should have a relationship linking them like “is translation of” each other. However, an expression in a target language that “is translation of” either a name of an entity expressed in a source language or an official

expression in an inter-lingual language is uni-directionally linked and hence cannot be traced the other way around. A frame for expressions in a target language should contain all possible translations from any available corpora in the target language. It is my aim to establish a triangulated terminology database in the Danish e-government domain based on an ontology-based terminology management system developed by Copenhagen Business School [4]. As the next step, I would like to investigate and to identify a base-line CLIR system that can integrate the triangulated terminology management approach.

FIGURE 2: Inconsistent pivot- and transitive translations

FIGURE 3: Ontology-based terminology management

FIGURE 4: Triangulated terminology management

REFERENCES.

- [1] Gollins, T. and Sanderson, M. (2001) Improving Cross Language Information Retrieval with Triangulated Translation, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States:pp.90-95
- [2] Ballesteros, L. (2001) Cross Language Retrieval via transitive translation, In Croft W. B. (ed). *Advances in Information Retrieval: recent Research from the CIIR*, Kluwer Academic Publishers, pp.203-234
- [3] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004) WordNet::Similarity – Measuring the Relatedness of Concepts. Available from: <http://search.cpan.org/dist/WordNet-Similarity>.
- [4] Madsen, B, Thomsen, H. and Wenzel, A (2006) i-Term for NORDTERM 5th *International Conference on Language Resources and Evaluation (LREC 2006)*, Workshops Proceedings: W16 Terminology Design: Quality Criteria and Evaluation Methods (TermEval). Genova, Italy