Automatic Translation of Multi-word Labels

Grzegorz Protaziuk, Marcin Kaczyński and Robert Bembenik

Abstract Application of semantic resources often requires linking phrases expressed in a natural language to formally defined notions. In case of ontologies lexical layers may be used for that purpose. In the paper we propose an automatic machine translation method for translating multi-word labels from lexical layers of domain ontologies. In the method we take advantage of Wikipedia and dictionaries services available on the Internet in order to provide translations of thematic texts from a given area of interest. Experimental evaluation shows usefulness of the proposed method in translating specialized thematic dictionaries.

Keywords Domain label translation · Automatic translation · Wikipedia application

1 Introduction

Semantic resources such as ontologies are becoming more and more important in various information systems nowadays. Domain ontologies in IT are considered to be formal descriptions of selected pieces of the world enabling various beings to share common understating of those areas of interest. The common understanding concerns a conceptual layer of ontologies defining semantics. However, in many situations where text written in a natural language is involved, associations between phrases expressing a certain notion in the natural language and a formal definition of that concept stored in an ontology are needed. Lexical layers of ontologies may be perceived as a kind of bridge [11] linking phrases in a natural language to an ontology concept, but usually a newly built ontology has a lexical layer expressed

G. Protaziuk (✉) · M. Kaczyński · R. Bembenik
Institute of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warszawa, Poland
e-mail: G.Protaziuk@ii.pw.edu.pl

M. Kaczyński
e-mail: M.Kaczynski@elka.pw.edu.pl

R. Bembenik
e-mail: R.Bembenik@ii.pw.edu.pl

D. Ryżko et al. (eds.), *Machine Intelligence and Big Data in Industry*,
Studies in Big Data 19, DOI 10.1007/978-3-319-30315-4_9

only in one or few languages. The practical usefulness of such an ontology is more or less limited. Manual translation is a costly and time consuming process. Also, such translation may require involvement of a domain specialist as it is quite probable that a domain ontology includes definitions of very specialized concepts.

This paper addresses that problem and proposes an approach to automatic generation of possible translations of labels included in a lexical layer of an ontology. The introduced method uses external resources available on the Internet and returns phrases taken from thematic documents as the proposed translations.

The problem of translating lexical layers of ontologies is related to the general problem of automatic machine translation, particularly translation of dictionaries. Source data for algorithms of automatic translation can be perceived twofold, as (1) a multilingual comparable corpus, a set of collections of texts referring to similar topics, and (2) a parallel corpus, a set of collections of linked texts being adjusted translations. In methods dealing with parallel bilingual corpora [1, 6, 8, 13] usually firstly pairs of corresponding sentences from texts are identified, and then pairs of corresponding words/phrases are determined for the sentences paired in the previous step. Typically, such an adjustment is performed by means of a probabilistic model of languages proposed by authors of a translation method. Examples of application of statistical machine translations methods can be found e.g. in [4]. Also, a lot of methods have been introduced in the literature, in which a statistical model is enhanced by various linguistic information or semantics. For example, in [3] the source-context similarity approach to creation of a statistical model was applied. In that method a translation unit consists of three elements: phrase-source-side, phrase-target-side, and source-context (a source sentence from which a translation unit was originally extracted). The source-context is used for calculation of similarity between an input sentence to be translated and a translation unit. In [2] the synchronous context-free grammars (SCFGs) for creating statistical models of translation were used. Authors incorporated modality and named entity tags as higher-order symbols in translations rules. A review of such methods can be found in [5, 14]. Methods of statistical machine translation can provide very high precision of translations but (1) require parallel corpora that are rather rare and (2) taking into consideration changes in languages requires rebuilding of language models.

In methods dealing with comparable corpora it is usually assumed that words often occurring together in texts in a source language also often occur together in texts in a target language. Such relation between two words A and B takes place if a distance (number of words) between A and B in a given text is not greater than a certain threshold. In these methods also a base dictionary including translations of selected words is applied for comparing contexts of words and for building a resultant dictionary. In [7, 12] such dictionary is an input parameter, whereas in [9, 10] it is created from scratch. A dictionary is built based foremost on identical spelling of some words in both languages and usage of rules which eliminate typical differences in spelling of words in the considered languages. The reported results obtained by means of such methods were essentially worse than from methods using parallel corpora. However, they may be applied in many more situations because comparable corpora are significantly more common (e.g. Wikipedia) than the parallel ones.

The rest of the paper is structured as follows. Section 2 presents our approach to labels translation, Sect. 3 contains a formal description of the proposed method, Sect. 4 discusses results of the experiments, and, finally, Sect. 5 concludes the paper.

2 The Translation Method

The proposed method of translation is aimed at translating multiword labels from a lexical layer of a domain ontology. In our approach the focus is set on finding correct translations coming from vocabulary used in a given thematic area. We search for translations in a text document repository and assume that proper translations of labels occur in those texts. In our method external services are used to obtain data applied for building proper translations. We use data from Wikipedia both for verification of correctness of the prepared translations as well as for selecting the most suitable translations for a given domain of interest. We use Wikipedia also for finding translations.

We begin the presentation of a proposed approach from a short description of Wikipedia from the perspective of our method and then we present in detail the proposed translation procedure.

2.1 Wikipedia

Wikipedia is an online, publicly accessible, multilingual encyclopedia which accumulates knowledge of many different fields. It has a multi-level structure of references and categories. The categories may be perceived as the domain identifiers and a hierarchy of categories may be used for selecting pages from a given domain (more or less general). We use categories for building a domain-specific repository of articles. In Wikipedia there are three types of links, namely: internal, external, and inter-lingual. The last ones indicate links connecting articles written in different languages describing the same notions. This makes it possible to use these links for finding correct translations. Moreover, it is easy to determine categories (domains) associated with inter-lingual links which, in turn, allows checking if a found translation refers to a given thematic area. In the method we also take advantage of disambiguation pages grouping synonyms and the redirection mechanism used in Wikipedia to represent the issues that occur with several different names.

2.2 Translation Procedure

The proposed method of label translations consists of the three following ways of translation:

1. *Translation based on Wikipedia* in which we take advantage of inter-lingual links and the redirection mechanism.
2. Translation utilizing *the TransLab* procedure, which consists of the following steps:

   - translations of single words—obtaining possible translations for each word included in a label;
   - generation of candidate translations of multi-word labels;
   - verification of the candidate translations.

3. Translation based on translator services

   - generating candidate translations based on Internet translator services;
   - verification of candidate translations.

Repository. A text document repository used in the translation method may be any set of thematic documents written in a target language. It is assumed that translated labels and documents in a repository concern the same domain area. In our implementation texts are retrieved from Wikipedia based on keywords (which are equated with names of Wikipedia categories) provided by a user. We use the list of pages associated with Wikipedia categories to obtain the needed articles. Also, we recursively retrieve articles linked with subcategories of the previously searched categories.

Translation based on Wikipedia. In this method we search for pages whose titles are the same as the translated label. If such a page (or page found by means of the Wikipedia redirection mechanism) has an inter-lingual link to a page written in a target language the title of a linked page is treated as a translation. As in our research we focused on translating labels from English into Polish and the most English titles of pages in Wikipedia are composed of words in their singular forms, we additionally applied the following transformations of words included in the labels:

- removing suffix 's', e.g. computers → computer;
- changing suffix 'ies' into 'y', e.g. cities → city;
- changing suffix 'ves' into 'f' and 'fe', e.g. (halves → half, knives → knife),
- changing suffix 'es' into 'e' and removing that suffix, e.g.: bridges → bridge, dresses → dress.

As the applied method of transformation is very simple and it does not take into account the numerous exceptions that occur in English it does not guarantee generation of the correct singular forms of words. Generally, usage of that method should give good results, but we cannot exclude situations in which usage of the method may result in incorrect translations. As we build a domain repository from Wikipedia pages we can perform a simple verification of the found translations, namely: a given translation is valid only if it is taken from a page included in the domain repository.

Translation of single words. The possible translations of each word included in a label are retrieved by means of machine translation services available on the Internet. For this purpose a translated label is split into single words, and with each such word

a list of the found translations is associated. In our experiments we used the following services: *Bing Translator* (www.bing.com/translator), *Google Translate* (http://google.translate.pl), *1 A Tradovium* (http://www.1a.biz.pl/slownik), and Translate.pl (http://translate.pl).

Generation of candidate translations. For generating candidate translations for labels single words are used, namely: for a label consisting of n words all possible sequences of terms $t_1, t_2, \ldots, t_n$ are created, where term t_k belongs to the list of translations associated with the word at the kth position in the translated label. As the order of words in a proper translation of a label in a target language may be different from the order of words in the translated label we generate candidate translations for all permutations of words included the translated label.

Translation based on translator services. For research purposes we implemented a method for obtaining translations of whole phrases from the Google Translate service. The main disadvantage of the service is that it returns only one version of the translation. Also, it should be noted that the service returns a translation for any label, and in the worst case a resultant translation is exactly the same as a source phrase.

Verification of the candidate translations. The aim of this step is to select translations from the set of candidate translations that are valid in a target language and are appropriate for a given thematic area. For that purpose the domain repository is used. Generally, we check whether a candidate translation occurs in text documents contained in that base; if so the translation is considered correct. Verification is performed for both methods of generating candidate translations: *TransLab* procedure and *Translation based on translator services.*

2.3 Searching Phrases in a Domain Repository

Determining occurrence of a word in texts

A proper translation of a multi-word label may include words that are not in their basic form. It especially concerns situations in which a target language is highly inflected (e.g. Latin, Polish). It is probable that by using machine translation services we do not obtain all possible grammatical forms of translations of a given word. Usually because of that such services return translations only in the basic word forms. A variety of grammatical forms of a given word is a cause of a problem with proper determination of occurrences of that word in texts stored in a domain repository. In order to address that problem we defined a *represented relation* between words, namely: a word A included in a label is in a represented relation with a word Bincluded in a text if the word B may be considered an occurrence of the word A in that text. In the sequel, we denote that relation as $\Gamma(A, B)$. We used two methods for determining $\Gamma(A, B)$:

- By using dictionary variations (inflections, conjugations, etc.) of words. Such dictionary may be seen as a set of the following pairs: a word and a list of its all possible variations. In this method two words are in a represented relation if there is a list of variations that includes those both words. (In the experiments we used a dictionary available at the address: http://sjp.pl/slownik/odmiany/.)
- By computing similarity between words. The similarity between two words (character strings) a and b is calculated by means of $sim(a, b)$ function defined in the following way:

$$sim(a, b) = \frac{1.0}{1.0 + \frac{lev(a,b)}{min(|x|,|y|)}}$$

where: lev(a, b) is the Levenshtein distance between a and b,
$|x|$—number of characters in x

The similarity of two strings is a value form the range $\langle 0, 1\rangle$. The similarity is equal to 1 if two strings are the same.

Efficiency of searching in a domain repository The number of candidate translations may be very high, e.g. if a translated label consists of three words and each word is translated into five terms and each term has 4 variants, the number of candidate translations is equal to $(5^*4)^3 = 8000$. A verification of candidate translations may be performed in an efficient manner provided that a repository management system offers fast access to sentences in which a given word occurs.

3 Formal Description of the TransLab Procedure

In the sequel we use the following notation:

- W—a set of words.
- $W^{\text{lang}}/T^{\text{lang}}$—a set of words/terms from a given natural language *lang*.
- λ—label. A label is a sequence of words from a given finite set of words. $\lambda = \langle w_1, w_2, \ldots, w_n\rangle$, $w_i \in W$. A label built from k words is a *k-label*.

3.1 Translation of a Single Word

Function $wordtran_{A\rightarrow B}$ is a translation function which for a given word w from a language A returns a set of terms (possible translations of the word w) from a language B. Formally:

$$wordtran_{A\rightarrow B}: w^A \rightarrow T^B,$$

where w^{A}—a word from a language A, $\mathrm{w}^{\mathrm{A}} \in \mathrm{W}_{\mathrm{langA}}$;
and T^{B} is a set of terms from a language B $T^B = \{t_1, t_2, \ldots, t_n\}$, $t_i \in W_{\mathrm{langB}}$.

3.2 Generation of the Candidate Translations

Function $labeltran_{A\to B}$ is a translation function which for a given label consisting of words from a language A returns a set of sets of possible translations for each word included in the label. Formally:

$$labeltran_{\mathrm{A}\to\mathrm{B}}: \lambda^{\mathrm{A}} \to \cup_{i=1}^{n} \mathrm{T}_i^B$$

where λ^{A}—a label consisting of $n$words from a language A, $\lambda^{\mathrm{A}} = \langle w_{1,} w_2, \ldots, w_n\rangle$, $w_i \in W_{langA}$
T_i^B—a set of terms from a language B, which are possible translations of a word $\mathrm{W_i}$; $\mathrm{T}_i^B = wordtran_{\mathrm{A}\to\mathrm{B}}(\mathrm{w_i})$;
$\cup_{i=1}^{n} \mathrm{T}_i^B$—a set of sets T_i^B.

Example Given the label: *binary numeral system*

$wordtran_{\mathrm{E}\to\mathrm{P}}$ (binary) = {dwójkowy; podwójny; dwuskładnikowy; złożony z dwóch pierwiastków; dwuczłonowy}
$wordtran_{\mathrm{E}\to\mathrm{P}}$ (numeral) = {liczebnik; liczba; liczbowy; cyfrowy}
$wordtran_{\mathrm{E}\to\mathrm{P}}$ (system) = {system; układ; sieć; ustrój; reżim; metoda; organizm; po-rzadek; organizacja; systematyczność; formacja}
$labeltran_{\mathrm{E}\to\mathrm{P}}$ (binary numeral system) = {dwójkowy; podwójny; dwuskładnikowy; złożony z dwóch pierwiastków; dwuczłonowy}, {liczebnik; liczba; liczbowy;
cyfrowy},{system; układ; sieć; ustrój; reżim; metoda; organizm; po-rzadek; organizacja; systematyczność; formacja}}

For a given sequence $seqK(T^B)$ of sets T^B a candidate translation of a label $\lambda^A = \langle w_{1,} w_{2,} \ldots, w_k\rangle$ is a sequence of terms $ct = \langle t^1, t^2, \ldots, t^k\rangle$, where $t^j \in \mathrm{T}_i^B$. As term t^i is a sequence of words a candidate translation can also be presented as a sequence of words: $ct = \langle t^1\ t^2 \ldots\ t^k\rangle = \langle w_1^1, w_2^1, \ldots, w_{j1}^1, w_1^2, w_2^2, \ldots, w_{j2}^2, \ldots, w_1^k, w_2^k, \ldots, w_{jk}^k\rangle$, where w^i constitutes a term $t^i : t^i = \langle w_1^i, w_2^i, \ldots, w_{j1}^i\rangle$. A set of candidate translations, denoted as SCT_{seqK}, is a Cartesian product of the sets of possible translations of single words: $Sct_{seqK} = T_1^B x T_2^B x \ldots x T_K^B$.

As the order of words in a proper translation of a label in a target language may be different from order of words in the translated label the final set of candidate translations, denoted as Sct, is a sum of candidate translations for each possible order of sets T^B; $Sct = \cup_{l=1}^{k!} Sct_{seql}$, where k is a number of words in the translated label.

3.3 Verification of Candidate Translations

In the verification phase of the method we check if a considered candidate translation occurs in some sentences taken from documents stored in a thematic repository. Formally, a candidate translation $ct = \langle tw_1\ tw_2 \ldots\ tw_k \rangle$ occurs in a sentence $s = \langle w_1 w_2 \ldots w_n \rangle$ if there exist integers $i_1, i_2, \ldots, i_k$; $i_{j+1} = i_j + 1$, such that $\Gamma(tw_1, w_{i1})$, $\Gamma(tw_2, w_{i2})$, $\Gamma(tw_k, w_{ik})$.
Given:

- RD: a domain repository (in our case a set of articles);
- dT: a text document from the repository $dT \in RD$;
- s: a sentence (a sequence of words) $s = \langle w_1\ w_2 \ldots w_n \rangle$;

we define a candidate translation $ct = \langle tw_1 tw_2 \ldots\ tw_k \rangle$ to be a correct translation if $\exists s \subset dT$, $dT \in RD$ such that ct occurs in s. An adjusted sequence of words from the sentence s is added to the set of final translations. Also, for each proposed translation we provide basic statistics concerning the occurrence of that translation in the repository, namely:

- the number of occurrences in the repository (which is used to order the discovered translations);
- the number of articles containing that translation;
- the number of sentences containing that translation.

4 Experiments

In order to verify our assumptions we carried out a series of experiments. We used sets of labels which were taken from the specifications of the three ontologies presented below:

- CompSet—a set of 98 labels from the field of computer science selected in a random manner.
- ChemSet—a collection of 154 labels from the field of chemistry taken from the ontology available at http://ontology.dumontierlab.com/chemistry-primitive.
- ACMSet—a set of 1085 labels from the field of computer science and mathematics taken from the ontology ACM Computing Classification System available at http://totem.semedica.com/taxonomy/.

We evaluated the practical usefulness of the proposed procedure of translation by means of the precision measure, i.e. we calculated the percentage of correct translations in a resultant set of all translations. A translation ct was considered to be correct if it fulfilled one of the conditions of: (i) being equal to a reference translation rt, (ii) having the similarity between rt and ct greater than 0.8, (iii) for each word in ct

Table 1 Statistics summarizing the experiments per labels collections

	CompSet	ChemSet	ACMSet
Number of labels	98	154	1085
Labels translated correctly (first 5 proposals)	70.4 %	42.2 %	55.9 %
Labels translated correctly (first 10 proposals)	78.6 %	46.1 %	59.7 %
Labels translated correctly—all proposals	86.7 %	52.6 %	65.3 %
Labels not translated—all proposals	4.1 %	18.8 %	16.2 %
Avg. number of proposed translations	76	23	24
Avg. precision of proposed translations	0.26	0.22	0.41
Google Translator: correct translations	67.4 %	42.5	68.0
Bing Translator: correct translations	69.4 %	41.8	59.2

Table 2 Efficiency of the translation methods

Translation method	Labels set					
	CompSet		ChemSet		ACMSet	
	Found %	Correct %	Found %	Correct %	Found %	Correct %
Inter-lingual	46.9	91.3	17.5	100	29.4	99.0
Redirection	31.6	74.2	27.3	28.6	11.2	74.4
TransLab	91.8	82.2	76.6	61.7	52.7	56.3
Translators	84.7	80.7	61.0	69.2	38.2	70.5

having the similarity between words at the same position in *ct* and *rt* greater than 0.8. $sim(a, b)$ function was used to calculate the similarities.

The statistics summarizing the experiments are given in Table 1. The comparison of efficiency of different ways of translation is provided in Table 2. In this table the following notation is applied:

- *Inter-lingual*: translation based on the Wikipedia liter-ligual links;
- *Redirection*: translation based on the Wikipedia redirection mechanism;
- *Translab*: translation by means of the TransLab procedure;
- *Translators*: translation based on Internet translator services.

The achieved experimental results and their quality have been influenced not only by the quality of data returned by the used external dictionary or translator services, but also by the representativeness and quality of the documents stored in the domain repository. The complexity and the lengths of labels also had impact on the obtained results. In general, the longer and more complex labels, the worse the results.

The labels indicating concepts related to the field of computer science were composed of no more than four words and are commonly found in the texts. They are popular: about 47 % of these labels were translated using inter-lingual links. The labels of concepts related to the field of chemistry are much more complicated, which is reflected in the fact that less than 18 % of the translations were found by

means of inter-lingual links. The chemical ontology also includes labels indicating very specific notions, for which it was difficult in general to find appropriate Polish translations. Many labels in the ACM ontology are composed of multiple words. One label may be a composition of two other labels referring to different concepts. They are separated by comma or joined by a conjunction. The proposed method does not ensure obtaining good results in such situations. That is because such labels seldom if ever occur in thematic texts.

In general, the proposed method of translation allowed achieving better results than the results obtained by the means of Internet translators. One exception concerns the ACM ontology for which better results were obtained using the Google translator. It can be explained by the lack of occurrences of candidate translations in thematic documents.

5 Conclusions

In this paper we presented an approach to translating multi-word labels from a lexical layer of a domain ontology. The introduced method is focused on translations from English into Polish and some detailed solutions were adjusted to realize such translations. However, the applied schema of translation is a general idea and it may be adapted for the needs of translation between other languages. In our approach we find translations used in texts concerning the areas of interest by collecting data from an external dictionary and translation services and searching for the occurrence of the generated translations in documents stored in a thematic repository. Such an approach ensures that changes in languages will be reflected in the proposed translations. The conducted experiments showed practical usefulness of our method; we were able to find many correct translations. Additionally a user will receive information about the usage frequency of a given translation for the domain articles.

References

1. Brown P., Della Pietra S., Della Pietra V., Mercer R.: The mathematics of statistical machine translation: Parameter estimation. Comput. Linguist. (1993)
2. Baker K., Bloodgood M., Callison-Burch C., Dorr B.J., Filardo N.W., Levin L., Piatko C.: Semantically-informed syntactic machine translation: a tree-grafting approach arXiv (2014)
3. Banchs R.E., Marta R. Costa-jussà M.R.: A semantic feature for statistical machine translation. In: Proceedings of the fifth workshop on syntax, semantics and structure in statistical translation. Association for Computational Linguistics (2011)
4. Bojar O., Buck C., Federmann C., Haddow B., Koehn P., Monz C., Post M., Specia L. (eds.) In: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics (2014)
5. Costa-Jussà, M.R., Farrús, M.: Statistical machine translation enhancements through linguistic levels: a survey. ACM Comput. Surv. **46**(3), Article 42 (2014)

6. Gale W.A., Church K.W.: A program for aligning sentences in bilingual corpora. Comput. Linguist. (1993)
7. Haghighi, A., Liang, P., Berg-Kirkpatrick, T., Klein, D.: Learning bilingual lexicons from monolingual corpora. ACL. Associat. Comput. Linguist. (2008)
8. Kay M., Röscheisen M.: Text-translation alignment. Computat. Linguist. (1993)
9. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: Proceedings of ACL Workshop on Unsupervised Lexical Acquisition, ACL (2002)
10. Krajewski, R., Rybiński, H., Kozłowski, M.: A Seed Based Method for Dictionary Translation. Foundations of Intelligent Systems. LNAI. Springer, Berlin (2014)
11. Protaziuk, G., Wróblewska, A., Bembenik, R., Rybinski, H.: Lexical Ontology Layer—A Bridge between Text and Concepts. Foundations of Intelligent Systems. Springer, Berlin (2012)
12. Reinhard R.: Automatic identification of word translations from unrelated English and german corpora. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, ACL (1999)
13. Smadja, F., McKeown, K.R., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: a statistical approach. Comput. Linguist. (1996)
14. Wu D.: Toward machine translation with statistics and syntax and semantics, In: Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU'09) (2009)