Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova

University of Oslo, Department of Informatics
angelii@ifi.uio.no

Abstract

Machine-readable dictionaries play important role in the research area of computational linguistics. They gained popularity in such fields as machine translation and cross-language information extraction. Wiki-dictionaries differ dramatically from the traditional dictionaries: the recall of the basic terminology on the Mueller's dictionary was 7.42%. Machine translation experiments with the Wiki-dictionary incorporated into the training set resulted in the rather small, but statistically significant drop of the the quality of the translation compared to the experiment without the Wiki-dictionary. We supposed that the main reason was domain difference between the dictionary and the corpus and got some evidence that on the test set collected from Wikipedia articles the model with incorporated dictionary performed better.

Keywords: machine-readable bilingual dictionary, Wiki-dictionary, statistical machine translation

1. Introduction

Machine-readable bilingual dictionaries are employed in fields such as machine translation and cross-language information extraction. Possibilities for automatic generation of high quality resources of this type are being actively investigated by the research community because manual development is expensive and time-consuming. The main challenges for this task are found in achieving a reasonable level of accuracy, excluding noisy data and providing required coverage of terminology. With efficient methods for creation of bilingual dictionaries for different domains, we can, for example, experiment with usage of these dictionaries in the alignment modules of translation systems.

In this article we investigate the quality and the content of an English-Russian dictionary (Wiki-dictionary)[1] created from Wikipedia. In order to perform an in-depth evaluation of the resulting dictionary, we did named entity recognition and classification, computed the recall of the translation pairs on the traditional English-Russian Mueller's dictionary, collected corpus statistics from ÚFAL Multilingual Corpora[2] and incorporated the dictionary into a statistical machine translation system.

Even though it has been repeatedly shown that Wiki-dictionaries have many advantages, our experiments with the Wiki-dictionary show that it is important to clearly understand the domain to which they are applicable, otherwise improper usage may lead to drop of accuracy in the translation task.

2. Related Work

In the last decade the online encyclopedia Wikipedia has gained popularity because it is a multilingual, dynamic and rapidly growing resource with user-generated content. Wikipedia link structure was exploited, for example, for linking ontology concepts to their realizations in text (Reiter et al., 2008), for generating comparable corpora using a link-based bilingual lexicon for identification of similar sentences (Adafre and de Rijke, 2006).

(Erdmann et al., 2008) propose a method for creating a bilingual dictionary from interlanguage links, redirect pages and link texts. The number of backward links of a page is used to estimate the accuracy of a translation candidate because redirect pages with wrong titles or titles that are not related to the target page usually have a small number of backward links. The authors show the advantages of their approach compared to dictionary extraction from parallel corpora and manual crafting.

(Rohit Bharadwaj G, 2010) discuss the iterative process of mining dictionaries from Wikipedia for under-resourced languages, though their system is language-independent. In each step near comparable corpora are collected from Wikipedia article titles, infobox information, categories, article text and dictionaries built at previous phases.

(Yu and Tsujii, 2009) automatically extract bilingual dictionary from Chinese-English comparable corpora which is build using Wikipedia inter-language links. Single-noun translation candidates for the dictionary are selected by employing context heterogeneity similarity (a feature that claims that the context heterogeneity of a given domain-specific word is more similar to that of its translation in another language than that of an unrelated word in the other language) and then ranked with respect to dependency heterogeneity similarity (a feature that assumes that a word and its translation share similar modifiers and head).

There has also been research done on the effectiveness of the usage of bilingual dictionaries in machine translation. A bilingual dictionary can be used as an additional knowledge source for training of the alignment models. The parameters of the alignment models can be estimated by applying the EM algorithm. A dictionary is assumed to be a list of word strings (e, f) where e and f can be single words or phrases.

One of such methods of integrating of the dictionary into EM algorithm, described in (Brown et al., 1993), requires adding every dictionary entry (e, f) to the training corpus with an entry-specific count called effective multiplic-

[1] `http://folk.uio.no/angelii/wiki_dic.htm`

[2] `http://ufal.mff.cuni.cz/umc/cer/`

ity $\mu(e, f)$. Results of experiments in (Brown et al., 1993) showed that the dictionary helps to improve the fertility probabilities for rare words.

Another method described in (Och and Ney, 2000) suggests that effective multiplicity of a dictionary entry should be set to a large number if the lexicon entry occurs in at least one of the sentence pairs of the bilingual corpus and to low value if it doesn't occur in the corpus. The approach helps to avoid a deterioration of the alignment as a result of a out-of-domain dictionary entries.

3. Method

We created the Wiki-dictionary using the interlanguage links and redirect pages methods described in (Erdmann et al., 2008) (see Figure 1 for more details). The first assumption is that the titles of the articles connected by the interlanguage link are translations of each other. The second assumption is that the titles of redirect pages are the synonyms of the title of the target page. We collected titles of the articles conjoined by the interlanguage links and redirects from Wikipedia and created the dictionary from them.

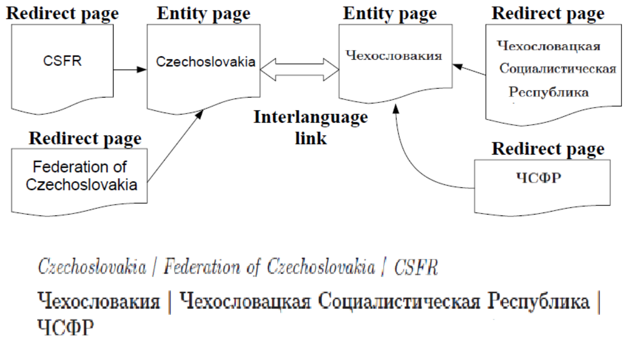
Redirect page
Entity page
Entity page
Redirect page
CSFR
Czechoslovakia
Чехословакия
Чехословацкая Социалистическая Республика
Interlanguage link
Redirect page
Federation of Czechoslovakia
Redirect page
ЧСФР
Czechoslovakia | Federation of Czechoslovakia | CSFR
Чехословакия | Чехословацкая Социалистическая Республика | ЧСФР

Figure 1: The interlanguage links and redirect pages methods for the Wiki-dictionary development

We included in the dictionary the Russian-English translation pairs that are present in the Russian Wikipedia dump and are absent from the English Wikipedia dump and the English-Russian translation pairs that are present in the English Wikipedia dump and are absent from the Russian Wikipedia dump. We have such data because of two reasons: first, the dumps were made on different dates, during this gap Wikipedia editors made changes to the encyclopedia, second, some articles have only one-way mappings, e.g. there is an interlanguage link from Russian article to English article but there is no interlanguage link from this English article or any of its redirect pages to the given Russian article. For example, Russian article "Случайные знаки" has an interlanguage link to the English article *"Accidental (music)"*. The latter article has a bi-directional interlanguage link with the article "Альтерация (музыка)" which means it is not connected with the article "Случайные знаки" in English-Russian direction.

4. Evaluation

In order to estimate the proportion of named entities in the Wiki-dictionary, we used the heuristics suggested in (Bunescu and Pasca, 2006) and some additional heuristics (e.g. a one-word title is a named entity if it contains at least one capital letter and at least one digit). The numbers show that 88% of the translation pairs are named entities while only 12% are non-named entities (non-NEs). For comparison, only 7.5% of entries in the traditional Mueller's dictionary contain named entities.

Having a large percentage of named entities in the Wiki-dictionary, it was interesting to see the distribution of classes of named entities. We performed named entity recognition and classification in order to learn more about the content of the dictionary. Using Wikipedia's own category system we labeled the Wiki-dictionary with the standard named entity tags (PER, LOC, ORG, MISC) which can be further used by the information extraction tools. We implemented a bootstrapping algorithm for the named entity classification task (Knopp, 2010). Each named entity class is represented as a vector of Wikipedia categories and the algorithm computes similarity between the category vectors of unclassified articles and the named entity class-vectors in each iteration. The class with the highest similarity score is assigned to the corresponding articles and the categories of these new classified articles are added to the vectors of their respective named entity class.

We manually marked-up a random sample of 300 dictionary entries and found out that the results of the automatic named entity recognition had an accuracy rate of 76.67% and the true distribution of the classes on the sample was:

- 24.33% entities of class PER;
- 2.67% entities of class ORG;
- 29.33% entities of class LOC;
- 15.67% entities of class MISC;
- 72% named entities in total.

In order to evaluate the Wiki-dictionary we checked whether Wiki-dictionary covers the vocabulary of the unidirectional English-Russian dictionary by V. K. Mueller. We obtained a machine readable version of the Mueller's dictionary in four plain text files: abbreviations, geographical names, names and base dictionary. The size of the Muller's dictionary is shown in the Table 1 ("Names" is a list of personal names, "Base" is a list of translation pairs that are non-NE). The Wiki-dictionary contains 348,405 entries.

The algorithm works the following way. It searches for the exact match of the lowercased English word from the Mueller's dictionary in the Wiki-dictionary, e.g. we take a record

`Czechoslovakia`
_ист. Чехословакия
Transliteration: _ist. čexoslovakija

from the Mueller's dictionary and search for the word *"Czechoslovakia"* in the Wiki-dictionary. If the entry of the Wiki-dictionary with this word is found, we collect all the Russian translations from the Wiki-dictionary. In our example the corresponding Wiki-dictionary record would

Mueller's dictionary file	**Geographical names**	**Names**	**Abbreviations**	**Base**
Number of entries	1,282	630	2,204	50,695

Table 1: Size of Mueller's dictionary files

Mueller's dictionary file	**Geographical names**	**Names**	**Abbreviations**	**Base**
Recall of the Wiki-dictionary	82.18%	75.88%	22.64%	7.42%

Table 2: Recall of the Wiki-dictionary on the Mueller's dictionary

be (the entry is shortened):

Czechoslovakia | Federation of Czechoslovakia | Czechoslowakia | Czechaslavakia | CSFR
Чехословакия | Чехословацкая Социалистическая Республика | Чешско-Словацкая Социалистическая Республика | Чешско-Словацкая Федеративная Республика | ЧСФР
Transliteration: čexoslovakija | čexoslovackaja socialističeskaja respublika | češsko-slovackaja socialističeskaja respublika | češsko-slovackaja federativnaja respublika | čsfr

We concatenate all the lines of the translation part in the Mueller's dictionary in one line and for each translation from the Wiki-dictionary we check if it occurs as a substring in Mueller's dictionary translation.

The reason why we concatenate the translation part in one line and search the Wiki-dictionary translations as substrings, is that the Mueller's dictionary often provides an explanation of a term rather than just a simple translation.

The results of the evaluation are summarized in Table 2. The highest recall we obtain is for the geographical names, 82.18%, while for the names we have 75.88%. Surprisingly, the highest recall we have obtained for the abbreviations, even taking the English expansions of the abbreviations into the account, is only 22.64%. Recall for the base dictionary is only 7.42% which shows the low coverage of non-NEs in the Wiki-dictionary.

Words that are included in Wikipedia but not in the Mueller's dictionary are largely (a) very specific terms (such as *monogeneric*, *apature*, *rem sleep parasomnias*, *tokamak*, *tropidoboa*) that are more likely to be present in field-specific dictionaries rather than in general lexicon and (b) particular named entities (local geographical names such as *Santana do Acaraú*, *Lake Semionovskoye*, *Emelyaianovski district*; names of public people such as *Edvard Speleers*, *Princess Theresa of Bavaria*, *Alberto Medina Briseno*, *William de Lyon*; football teams such as *FC Zauralets Kurgan*; car models such as *Mercedes-Benz W221*; etc.).

5. Machine Translation Experiments

For the machine translation experiments we used sentence-aligned ÚFAL Multilingual Corpora (UMC) and we chose the Moses[3] toolkit which is a complete machine translation system for academic research. UMC is a parallel corpus of texts in Czech, Russian and English languages created for the purpose of machine translation. The source of the content are news articles and commentaries from The Project Syndicate[4].

We were interested in the frequency of dictionary phrases in corpus data and we had a goal to do pre-evaluation of the corpus to find out whether we could use it for machine translation experiments with the dictionary. We therefore collected statistics of occurrences of the translation pairs from the Wiki-dictionary in the UMC. The evaluation was done by word forms (using a tokenized version of the dictionary) and by normal forms (using a tokenized lemmatized version of the dictionary and a normalized version of the corpus data). Results show that translation pairs from the Wiki-dictionary are present in the corpus but not to a large extent. Approximately 28% of the non-normalized sentence pairs from the training set don't contain any translation pairs from the Wiki-dictionary, while approximately 24.7% of the non-normalized training set contains exactly one translation pair from the Wiki-dictionary.

First, we performed several experiments without the Wiki-dictionary and achieved the highest BLEU score of 24.76 using the English monolingual data from Europarl corpus[5] as additional data for training a language model.

We then incorporated the Wiki-dictionary into the training set: the dictionary was split into pairs of synonyms and appended to the end of the UMC training set. The inclusion of a dictionary as an additional parallel corpus data is the standard method. But this resulted in a drop of BLEU score, the best value we got was 20.42.

We used paired bootstrap re-sampling to estimate the statistical significance of the the difference in BLEU score between the model created with and without the Wiki-dictionary. As the difference between BLEU scores of the systems was small, we couldn't be sure if we could trust automatic evaluation results that one system outperformed the other on the test set. Our question was if the difference in test scores was statistically significant.

The approach is described in (Koehn, 2004). We collected 1000 trial sets of the size 300 sentences from the original test set (which had the size of 1000 sentences) by random sampling with replacement. We computed BLEU score for both systems in question on each of the 1000 trial sets and calculated how many times one system outperformed the other.

[3] `http://www.statmt.org/moses/`

[4] `http://www.project-syndicate.org/`

[5] `http://www.statmt.org/europarl/`

We compared the models that were created without an additional corpus for language model training. The results are summarized in the Table 3. According to our evaluation, 3-gram model without the Wiki-dictionary is better than the model trained with the Wiki-dictionary with 98.5% statistical significance, 4-gram model is better with 96% statistical significance and 5-gram model is better with 87.1% statistical significance.

A possible explanation for the drop is the domain difference of the corpus and the Wiki-dictionary. UMC corpus contains texts from the collection of the news articles and commentaries from a single resource The Project Syndicate while Wikipedia is an Internet encyclopedia. Typically, the more data is used for the translation model training the higher translation performance can be achieved. However, the significant amount of out-of-domain data added to the training set cause the drop of the translation quality (Hildebrand et al., 2005). In such a case a general translation model that was trained on in-domain and out-of-domain data does not fit the topic or style of individual texts. For the ambiguous words the translation highly depends on the topic and context they are used in.

The UMC training set contained a significant number of sentences that comprised zero or only one word from the Wiki-dictionary. We believe that might mean that the domains of the Wiki-dictionary and the UMC corpus are quite different. We suppose that was the reason of the lower quality of the translation that we got from the model trained on the train set with the Wiki-dictionary incorporated in it.

Therefore we collected a new test set using the text of three articles from Wikipedia (Wiki-set). The text of the articles needed pre-processing. First, we converted MediaWiki text into plain text using the Java Wikipedia API (Bliki engine)[6] which is a parser library for converting Wikipedia wikitext notation to other formats. The class PlainTextConverter from this library can convert simple Mediawiki texts to plain text. Secondly, we removed that traces of template markup (e. g. {{cite web}}) that still remained after removing Mediawiki markup. Thirdly, we split the text into sentences with the script split-sentences.perl written by Philipp Koehn and Josh Schroeder as part of Europarl v6 Preprocessing Tools suit[7]. The tool uses punctuation and capitalization clues to split paragraphs of sentences into files with one sentence per line. Fourthly, we performed tokenization using the same script as in Chapter 2, the script tokenizer.perl from Europarl v6 Preprocessing Tools suit. Finally, we corrected the automatic tools errors and removed the remaining noise manually.

Both the UMC test set and the Wiki-set consist of 1000 sentences, but there are 22,498 tokens in the Wiki-set while the UMC test set contains 19,019 tokens. Since there is no gold standard, we manually compared the quality of the translations produced by the models trained with and without the Wiki-dictionary on two random samples of 100 sentences collected from the UMC test set and from the Wiki-set. Table 4 presents the results of this manual ranking. In most of the cases one of the systems was ranked higher than the other because of the better representation of the meaning of the original sentence. In many other cases the missing words and grammatical structure played the key role in the final decision. There were several pairs for which one translation was preferred against the other because of the vocabulary, as some synonyms suit particular contexts better than the other synonyms. The model trained without the Wiki-dictionary performs better on the sample from the UMC test set; it is ranked first on 55 sentences. This outcome corresponds to the BLEU evaluation results. The model trained with the Wiki-dictionary is ranked first on 50 sentences of the sample from the Wiki-set while the outputs of the two models are of indistinguishable quality on 6 sentences. This brings some evidence that the Wiki-dictionary can be useful when it is applied to the appropriate domain.

Out-of-vocabulary (OOV) words are the words of the source language that the machine translation system didn't manage to translate into the target language. The total number of OOV words is less for the model trained with the Wiki-dictionary on both test sets. As we expected there are many cases when the model trained without the Wiki-dictionary didn't translate named entities while the model trained with the Wiki-dictionary recognized and translated the named entities correctly. For example,

```
<s1>sociologist at yale university
immanuel валлерстайн believes that
by 2050 , lenin inevitably become a
national hero russia .  </s1>
<s2>marketing sociology at yale
university , immanuel wallerstein
believes that by 2050 , lenin
inevitably will be the national hero
russia .  </s2>
```

The number of OOV words is twice bigger on the Wiki-set while the sizes of the test sets are comparable. The increase in the number of OOV words is most likely caused by the shift of the topic.

6. Conclusions

In this work we evaluated a bilingual bidirectional English-Russian dictionary created from Wikipedia article titles. This dictionary is very different from the traditional Mueller's dictionary, e.g. most of the phrases and words are named entities, the recall of the common terminology is only 7.42% and at least 96% of the basic terminology that the Wiki-dictionary shares with the Mueller's English-Russian dictionary are noun phrases. Evaluation on the parallel ÚFAL Multilingual Corpora revealed that even though the translation pairs from the Wiki-dictionary occur in the corpus, there is a significant number of sentences (about 28%) that don't contain any terms from the Wiki-dictionary. Such statistics indicates that the dictionary doesn't properly cover the domain of this corpus. As a next step, we incorporated the Wiki-dictionary into a translation system. According to the BLEU score, paired bootstrapping, OOV words analysis and manual evaluation, the translation accuracy dropped compared with the models trained without the Wiki-dictionary. The difference in the domain of the corpus and the dictionary could explain this result. We got

[6] http://code.google.com/p/gwtwiki/

[7] https://victorio.uit.no/langtech/trunk/tools/alignment-tools/europarl/

Model 1	**Model 2**	**Statistical significance that model 1 outperforms model 2**
3-gram	3-gram + Wiki-dict.	98.5%
4-gram	4-gram + Wiki-dict.	96%
5-gram	5-gram + Wiki-dict.	87.1%

Table 3: The results of the paired boostrap re-sampling show the statistical significance of the fact that the models trained without the Wiki-dictionary outperform the models trained with the Wiki-dictionary

	Model **without** **Wiki-dict** is ranked first, # of sent.	Model **with** **Wiki-dict** is ranked first, # of sent.	Translations are equally bad/good, # of sent.
sample of 100 sent. from **UMC** test set	**55**	37	8
sample of 100 sent. from **Wiki-set**	44	**50**	6

Table 4: Manual ranking of the results

some evidence to support this hypothesis in the new experiment on the test set collected from Wikipedia. We found that the model trained with the Wiki-dictionary performed better than the model trained without the Wiki-dictionary according to OOV words analysis and manual evaluation.

7. References

S. F. Adafre and Maarten de Rijke. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer, and Surya Mohanty. 1993. But dictionaries are data too. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 202–205, Stroudsburg, PA, USA. Association for Computational Linguistics.

Razvan C. Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. An approach for extracting bilingual terminology from wikipedia. In *Proceedings of the 13th international conference on Database systems for advanced applications*, DASFAA, Berlin, Heidelberg. Springer-Verlag.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of EAMT*, Budapest, Hungary, May.

Johannes Knopp. 2010. Classification of named entities in a large multilingual resource using the Wikipedia category system. Master's thesis, University of Heidelberg.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 1086–1090, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nils Reiter, Matthias Hartung, and Anette Frank. 2008. A resource-poor approach for linking ontology classes to wikipedia articles. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 381–387, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vasudeva Varma Rohit Bharadwaj G, Niket Tandon. 2010. An iterative approach to extract dictionaries from wikipedia for under-resourced languages. ICON 2010, IIT Kharagpur, India.

K. Yu and J. Tsujii. 2009. Bilingual dictionary extraction from wikipedia. In *Proceedings of Machine Translation Summit XII*.