AUTOMATIC CREATION OF HUMAN-ORIENTED TRANSLATION DICTIONARIES

Antonova A. (antonova@yandex-team.ru),
Misyurev A. (misyurev@yandex-team.ru)

Yandex, Moscow, Russia

This paper addresses the issue of automatic acquisition of a human-oriented translation dictionary from a large parallel corpus. Automatically generated dictionary entries can enrich the output of a statistical machine translation system. We describe an automatic approach to the extraction of translation equivalents, and dictionary entry construction: grouping of synonymic translations, selection of illustrative context examples. The extraction of possible translations is based on statistical machine translation methods. The selection of lemmatized and linguistically motivated phrases is done with the help of morpho-syntactic analysis. In contrast to human-built dictionaries, an automatic dictionary usually contains a certain amount of noisy translations, as a consequence of systematic alignment mistakes and corpus imperfections. A noise reduction approach is proposed. We also provide the result of an evaluation experiment and the comparison of frequency distribution of words in the queries to the dictionary and the frequency distribution of words in plain text.

Keywords: parallel texts, bilingual dictionary extraction

1. Introduction

This paper describes an approach to the automatic construction of translation dictionaries. The approach is based on statistical machine translation (SMT) methods and can be applied to various language pairs. The dictionary entries are created automatically and contain translation variants grouped by meaning, reverse translations and context examples. For some languages, the dictionary entries can also include data prepared partly manually, e.g. transcriptions.

The automatic acquisition of translation equivalents from parallel texts has been extensively studied since the 1990s [7, 14]. The set of translation pairs is often referred to as bilingual lexicon. At the beginning, the automatically acquired lexicons served as internal resources for SMT [3], information retrieval (IR) [15], or computer-assisted lexicography [2, 4].

The growth of Internet and the current progress in search of web-based parallel documents [10, 12] makes it possible to automatically construct large-scale bilingual lexicons. Hence a new interesting possibility arises—to produce automatically acquired human-oriented translation dictionaries that have a practical application.

A machine translation system can output an automatically generated dictionary entry in response to all queries that are found in the dictionary. The percentage of short queries can be quite large, and the system benefits from showing several possible translations instead of a single result of machine translation (Fig. 1).

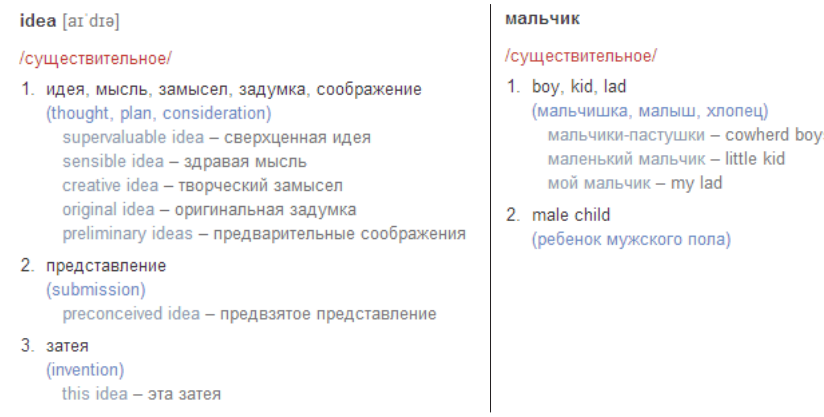


Fig. 1. Examples of dictionary entries in English—Russian and Russian—English dictionaries

The initial translation equivalents for an automatic dictionary bilingual lexicon can be extracted with the help of the techniques and tools developed for the phrase-table construction in SMT. The widely used word alignment and phrase extraction algorithms are described in [3, 9]. Though an SMT phrase-table actually consists of translation equivalents, it may differ substantially a human-oriented dictionary (Table 1). Additional algorithms are required to convert the initial translation equivalents into a dictionary.

Table 1. Differences between a human-oriented dictionary and an SMT phrase-table

Human-oriented dictionary	SMT phrase-table
Lemmatized entries are preferred.	Words and phrases in all forms are included.
Only linguistically motivated phrases are acceptable.	Any multiword combination is included.
Precision is important. Any noise is undesirable.	Having lots of low-probability noise is acceptable, since it is generally overridden by better translations.

The translation equivalents are organized into dictionary entries. The key of an entry is a word or phrase, usually lemmatized. Its translations are divided by the part of speech. Inside each part-of-speech class, the synonymic translations are grouped together. The groups are ordered according to their aggregate frequency.

Each group can be illustrated by reverse translations, and parallel context examples, drawn from the parallel corpus. Fig. 2 explains the structure of a dictionary entry for the word “French” in the English-Russian dictionary.

Fig. 2. The structure of a dictionary entry

Some aspects of the dictionary entry construction represent independent problems. The grouping of synonymic translations relies on the pre-constructed dictionary of synonyms, which also can be built automatically from the parallel corpus, as described in 2.5. The problem of selecting most illustrative context examples is discussed in 2.4.

In contrast to human-built dictionaries, an automatic dictionary usually contains a certain amount of noise—incomplete or totally incorrect translations. Yet, it may have some important advantages.

- Being objective and up-to-date. The frequency of uncommon or archaic translations is low. At the same time, the automatic approach often finds relevant translations, missed by a professional lexicographer [11].
- Improvement over time. With the possibility to process more parallel documents, the automatic dictionaries can potentially cover more words and phrases than the human-built dictionaries.
- Better flexibility. Since the procedure is fully automatic, it is easier to rearrange the dictionary, adjust its parameters (e.g. the precision/recall ratio, the maximum number of translations per a single entry). The translations can be ordered according to their frequencies or probabilities. This reduces the average time the user spends when looking for a particular meaning.
- Uniform approach to different language pairs.

The rest of the paper is organized as follows. We describe the overall system architecture in Section 2. We discuss the types of noisy translations and the noise detection approach in Section 3. The dictionary evaluation is described in Section 4. We conclude and discuss the applicability of the proposed approach to different language pairs in Section 5.

2. System Architecture

The overall process of the English-Russian dictionary construction is represented in Fig. 3.

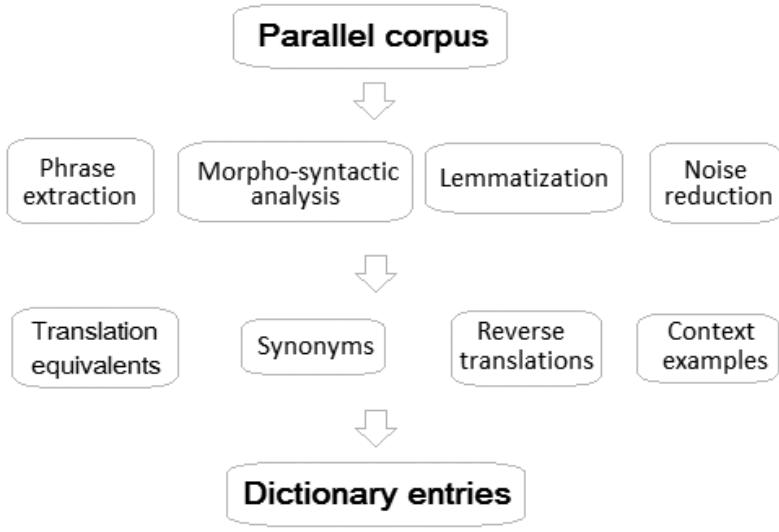


Fig. 3. The system architecture

2.1. Word Alignment, Morpho-Syntactic Analysis and Phrase Extraction

The parallel corpus is word-aligned and processed with English and Russian dependency parsers [1]. The initial phrase extraction is done as described in [8]. The maximum phrase length is limited by 3 words. We also discard the phrases where the words are not connected in any of the English and Russian parsing trees.

2.2. Lemmatization

The lemmatization is important; otherwise dictionary entries may contain many different forms of the same word, especially if the target language is morphologically

rich. The sentence preprocessing by an automatic lemmatization algorithm may introduce incorrect lemmas to the final dictionary. The reasons for that are ambiguity, heuristic lemmatization of unknown words, difficulties with the lemmatization of multi-word phrases. A possible way to overcome this problem is selecting a most lemma-like phrase pair among the real examples.

Each phrase pair can be assigned a key consisting of the lemmas of all words in it. The phrase pairs with the same key represent the translation equivalents in different forms. We select one best representative among them. The choice is made by taking into account the frequency of the unnormalized phrase pair and the morphological attributes.

2.3. Noise Reduction

Some undesired translation equivalents can be detected by simple heuristics. For example we remove the translations of punctuation, digits, or phrase pairs which occur extremely rarely. Still, there exist other types of noise which is more difficult to detect. We discuss this problem in detail in section 3.

2.4. Finding Context Examples and Reverse Translations to Illustrate the Meaning

Parallel context examples can help the user to distinguish the meaning of multiple translations. The examples are expected to be short well-formed grammatical phrases. We rely on the parsing trees to find such phrases. Besides, it is important to select not only the most frequent context example, but most distinctive and informative. This task is addressed by the metrics, such as mutual information, and the statistics of syntactic links between words.

Reverse translations also can be helpful to illustrate the word meaning.

2.5. Synonyms

The grouping of synonymic translations is done for two reasons. On the one hand, it visually shortens the list of translations and makes it easier to perceive. On the other hand, it also helps to differentiate faster between the different meanings.

The grouping procedure relies on the pre-constructed dictionary of synonyms, which is also built automatically from the same parallel corpus, from which the initial translation equivalents had been extracted. The idea behind automatic search of synonyms is that words with similar meaning are often translated by the same word in another language. We also used distributional similarity of syntactic contexts as an independent criterion of synonymity. Though each of the two factors—translation similarity and distributional similarity—could introduce incorrect synonyms, their intersection allows to increase the accuracy of the method.

2.6. Dictionary Construction for Different Language Pairs

The proposed procedure of dictionary construction requires morphological and parsing tools for both languages. Morphological tools are useful to determining possible parts of speech and lemmas of the word forms. Parsers are needed for the filtration of ungrammatical phrases and search of context examples. The part-of-speech disambiguation can be done within the parsing process, or with the help of a tagger. While the lack of such tools imposes certain restrictions on the dictionary content, the proposed approach can be still applicable in different cases.

If no parser is available, the dictionary will include neither multi-word translations, nor context examples. If morphology exists, but no morphological disambiguation tool is available, we can restrict the translations to those with identical part of speech.

3. Detection of Noisy Translations

The accuracy requirements are higher for a human-oriented dictionary, compared to a phrase-table used within an SMT system. The noise can appear as a consequence of systematic alignment mistakes and corpus imperfections, namely, nonparallel sentences, low-quality machine translation, language recognition mistakes. The following types of mistakes are common for automatically constructed dictionaries.

- Transliteration or translation by a word that belongs to a different language.
  челочка — chelochka

- Misspelled translation.
  monkey — обезъян

- Incomplete translation.
  доиграть — finish (finish playing)
  determined — определиться (be determined)
  present — памятный подарок (unforgettable present)

- Translation by an antonym. This can happen if one side incorporates a negation in its semantics, and the other sides has a negation as a separate word.
  eat — недоедать (be undernourished)
  unaware — подозревать (suspect)

- Translations of words with strict meaning. Though the highest-probability translations of proper names and colors are usually correct, some other variants may look unacceptable.
  yellow — белый (white)
  Russia — Украина (Ukraine)

- Translation by a common word.
  Russia — страна (country)

The most straightforward techniques of noise reduction in SMT phrase-tables is the filtration by frequency or probability thresholds [5]. However, in case of some systematic defects in the initial parallel corpus, a substantial amount of noise still survives, while many good translation equivalents are lost.

In addition to the translation probabilities, our approach to noise detection relies on the analysis of the parallel sentences in which a given translation pair occurred. There are several symptoms indicating that a sentence is a possible source of noisy translation:

- Unsafe one-to-one alignment. The intersection of HMM-based word alignments for two translation directions is a simple heuristic for finding the words that are confidently aligned to each other [9]. The percentage of such safe alignment points can vary in different sentences. However, its being too small is abnormal, and possibly indicates some defect to the given parallel sentences.
- High distortion of word order. Though some language pairs have different word order, the distance between the translations of subsequent input words is close to that of the input sentence. As well as unsafe one-to-one alignment, high distortion may indicate some defect to the given sentence pair.
- Bad syntactic structure. Some noisy translations originate from the sentences that seem to be an output of a poor-quality machine translation system. Bad translation often breaks up the syntactic structure of the output sentence.
- Many out-of-vocabulary words. Sentences containing many out-of-vocabulary words probably do not belong to the given language.
- Highly punctuated sentences. One can observe that sentences with lots of punctuation either are unnatural or contain enumeration. Large numeration lists are often not exactly parallel and can be aligned incorrectly, because many commas are mapped to each other.

4. Dictionary Evaluation

The evaluation of dictionary quality and the comparison of different dictionaries is a complicated task. Specifically, Tomaszczyk [13] considers multiple criteria for bilingual dictionary evaluation: equivalents, directionality, reversibility, alphabetization, retrievability, redundancy, coverage, currency, reliability. But these criteria are mostly qualitative and serve as a recommendation for a human expert reviewing a new dictionary.

One can also apply the standard information retrieval metrics, such as recall and precision. In this case, the manual gold standards must be prepared, which are difficult to construct, and are often biased towards the resource that the lexicographer consulted.

In this paper we evaluated the English-Russian dictionary against the following criteria: average number of translation variants in a dictionary entry, the percentage of incorrect translations and the percentage of extremely noisy translations. We used a manually annotated sample of translation equivalents randomly[1] drawn from the dictionary. The annotation task was to determine how well the given translation

[1] Random was used proportionally to the square root of joint frequency, in order to balance rare and frequent phrase pairs in the sample.

equivalent fits for a human-oriented translation dictionary. The annotators classified each translation according to the gradation represented in Table 2.

The average number of translation variants in a dictionary entry is 5.3 per query. A separate experiment has shown that a dictionary entry was found for 97 of 100 random queries to the dictionary.

The evaluation results show that 44.6% of the translation equivalents are unquestionably good, and 41.9% represent the words and phrases that were assessed as being redundant but not incorrect. For example, the dictionary includes many translations of trivial phrases as separate entries (*наша квартира—our apartment, our flat*). The total share of incorrect translations is about 13.5%.

Table 2. The percentage of translation equivalents in the English—Russian dictionary w.r.t. different quality types

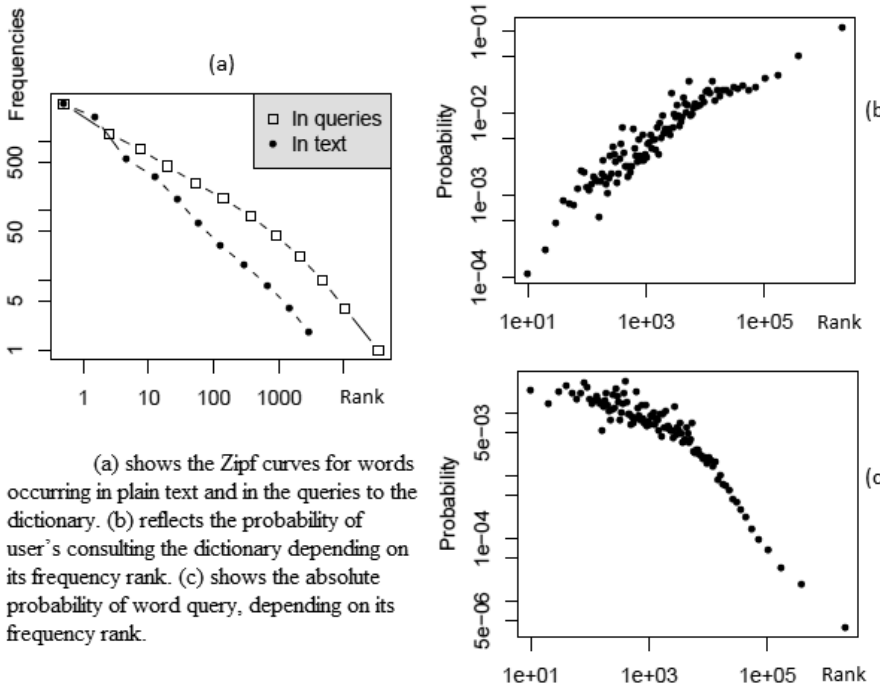
Type	Explanation	% in the dictionary
1	totally wrong or noisy (e.g. misspelled)	3.58
2	incorrect or incomplete translation	9.88
3	not a mistake, but unnecessary translation	41.90
4	good, but not vital	25.21
5	vital translation (must be present in human-built dictionary)	19.43

The evaluation does not take into account the frequency of queries and the order of translations in the dictionary entry. The first translations of frequent words are usually correct.

4.1. Analysis of Dictionary Use

The statistics of dictionary queries is relevant for the analysis of the dictionary quality and for the development of proper evaluation metrics. The properties of dictionary entries for frequent words may differ from those for rare words. Furthermore, the frequencies of words in dictionary queries may differ from their frequencies in text.

In this regard, we conducted an experiment, the purpose of which was to compare the frequency distribution of words in the queries to the dictionary with a uniform distribution, as well as to the frequency distribution of words in texts. We selected one-word queries from the dictionary log for a period of 18 days. Misspelled words were not considered. The frequencies of these words in queries were compared to the frequencies of the same words, collected over a large volume of texts on the Internet. The results are shown in Fig. 4. The points on the X axis correspond to words, ranked in descending order by frequency of occurrence in plain text. We can conclude that the distribution of words in dictionary queries is neither uniform, nor identical to the distribution of words in text. The Zipf curve for the queries decreases more slowly in the area of rare words. The reason is that rare words are likely to be unfamiliar, and the users need to consult the dictionary more often.



(a) shows the Zipf curves for words occurring in plain text and in the queries to the dictionary. (b) reflects the probability of user's consulting the dictionary depending on its frequency rank. (c) shows the absolute probability of word query, depending on its frequency rank.

Fig. 4. Comparison of the distributions of words in queries and text

Conclusion

We have described the procedure for the automatic construction of a large-scale translation dictionary, and the methods for augmenting dictionary entries with useful information, such as context examples, reverse meanings, grouping of synonyms. We also discussed the problem of detection of noisy translations, which is important for the human-oriented dictionary.

The results of the evaluation of the English-Russian dictionary demonstrated the perspectiveness of the overall approach, w.r.t. the coverage of the dictionary, and the depth of its entries. While the noisy translations still occur, their percentage is moderate. We provided the analysis of dictionary use, and discussed the difference between the distributions of words in dictionary queries and plain text.

The lemmatization and filtering ungrammatical phrases require additional morphological and syntactic tools. While the lack of such tools imposes certain restrictions on the dictionary content, the proposed approach is still applicable to many language pairs.

References

1. *Alexandra Antonova and Alexey Misyurev.* (2012). Russian dependency parser SyntAutom at the Dialogue-2012 parser evaluation task. Proceedings of the Dialogue-2012 International Conference.
2. *Sue Atkins.* (1994). A corpus-based dictionary. In: Oxford-Hachette French Dictionary (Introductory section). Oxford: Oxford University Press. xix—xxxii
3. *Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer.* (1993). The Mathematics of Statistical Machine Translation: Parameter estimation. Computational Linguistics, 19(2):263–312, June.
4. *Hartmann, R. R. K.* (1994). The use of parallel text corpora in the generation of translation equivalents for bilingual lexicography. In W. Martin, et al. (Eds.), Euralex 1994 Proceedings (pp. 291–297). Amsterdam:Vrije Universiteit.
5. *Philipp Koehn, Franz Josef Och, and Daniel Marcu.* (2003). Statistical phrase-based translation. In Proceedings of HLT-NAACL 2003, pages 127–133.
6. *Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondˇrej Bojar, Alexandra Constantin, and Evan Herbst.* (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pages 177–180, Prague, Czech Republic
7. *I. Dan Melamed.* (1996). Automatic construction of clean broad-coverage translation lexicons. In Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas, pages 125–134, Montreal , Canada
8. *Franz Josef Och and Hermann Ney.* (2000). Improved Statistical Alignment Models. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440–447, Hongkong, China.
9. *Franz Josef Och and Hermann Ney.* (2004). The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, vol. 30 (2004), pp. 417–449.
10. *Resnik, Philip and Noah A. Smith.* (2003). The web as a parallel corpus. Computational Linguistics, 29, pp. 349–380
11. *Serge Sharoff.* (2004). Harnessing the lawless: using comparable corpora to find translation equivalents. Journal of Applied Linguistics 1(3), 333–350.
12. *Jason Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch and Adam Lopez.* (2013). Dirt Cheap Web-Scale Parallel Text from the Common Crawl. To appear in Proceedings of ACL 2013.
13. *Tomaszczyk, J.* (1986). The Bilingual Dictionary under Review. Snell-Hornby, M. (Ed.). ZuriLEX'86 Proceedings. University of Zurich, Switzerland: 289–297.
14. *Dan Tufis¸ and Ana-Maria Barbu.* (2001). Computational Bilingual Lexicography: Automatic Extraction of Translation Dictionaries. In International Journal on Science and Technology of Information, Romanian Academy, ISSN 1453–8245, 4/3–4, pp. 325–352
15. *Velupillai, Sumithra, Martin Hassel, and Hercules Dalianis.* (2008). "Automatic Dictionary Construction and Identification of Parallel Text Pairs." Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS).