Improving the precision of automatically constructed human-oriented translation dictionaries

Alexandra Antonova
Yandex
16, Leo Tolstoy St., Moscow, Russia
antonova@yandex-team.ru

Alexey Misyurev
Yandex
16, Leo Tolstoy St., Moscow, Russia
misyurev@yandex-team.ru

Abstract

In this paper we address the problem of automatic acquisition of a human-oriented translation dictionary from a large-scale parallel corpus. The initial translation equivalents can be extracted with the help of the techniques and tools developed for the phrase-table construction in statistical machine translation. The acquired translation equivalents usually provide good lexicon coverage, but they also contain a large amount of noise. We propose a supervised learning algorithm for the detection of noisy translations, which takes into account the context and syntax features, averaged over the sentences in which a given phrase pair occurred. Across nine European language pairs the number of serious translation errors is reduced by 43.2%, compared to a baseline which uses only phrase-level statistics.

1 Introduction

The automatic acquisition of translation equivalents from parallel texts has been extensively studied since the 1990s. At the beginning, the acquired bilingual lexicons had much poorer quality as compared to the human-built translation dictionaries. The limited size of available parallel corpora often resulted in small coverage and the imperfections of alignment methods introduced a considerable amount of noisy translations. However, the automatimacally acquired lexicons served as internal resources for statistical machine translation (SMT) (Brown et al., 1993), information retrieval (IR) (McEvan et al., 2002; Velupillai, 2008), or computer-assisted lexicography (Atkins, 1994; Hartmann, 1994).

The current progress in search of web-based parallel documents (Resnik, 2003; Smith, 2013) makes it possible to automatically construct large-scale bilingual lexicons. These lexicons can already compare in coverage to the traditional translation dictionaries. 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 the short queries. 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 (Figure 1).

Figure 1: Examples of dictionary entries in two online statistical machine translation systems.

The initial translation equivalents for a 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 Brown et.al (1993) and Och (2004). Though an SMT phrase-table actually consists of translation equivalents, it may differ substantially from a traditional dictionary (Table 1).

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 58–66,
Gothenburg, Sweden, April 27, 2014.

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

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

While the problems of lemmatization and selection of linguistically motivated phrases can be addressed by applying appropriate morphological and syntactic tools, the problem of noise reduction is essential for the dictionary quality. The current progress in the automatic acquisition of similar Web documents in different languages (Resnik, 2003; Smith, 2013) allows to collect large-scale corpora. But the automatically found documents can be non-parallel, or contain spam, machine translation, language recognition mistakes, badly parsed HTML-markup. The noisy parallel sentences can be the source of lots of noisy translations — unrelated, misspelled, or belonging to a different language. For example, non-parallel sentences

> The apartment is at a height of 36 floors! (English)
>
> La plage est à 1 minute en voiture. (French: The beach is 1 minute by car.)

may produce a wrong translation "apartment - plage". Or, automatically translated sentences

> The figures in the foreground and background play off each other well. (English)
>
> Les chiffres du premier plan et jouer hors de l'autre bien. (French: The digits of the foreground and play out of the other well.)

may produce a wrong phrase translation "figures in the foreground - chiffres du premier plan".

An intuitive approach would be to apply noise filtering to the corpus, not to the lexicon. One could discard those sentences that deviate too much from the expected behavior. For example, sentences that have many unknown words and few symmetrically aligned words are unlikely to be really parallel. However, natural language demonstrates a great variability. A single sentence pair can deviate strongly from the expected behavior, and still contain some good translations. On the other hand, many noisy translations can still penetrate the lexicon, and further noise detection is necessary.

In a bilingual lexicon we want not just to lower the probabilities of noisy translations, but to remove them completely. This can be regarded as a binary classification task — the phrase pairs are to be classified into good and noisy ones.

Different types of information can be combined in a feature vector. We take advantage of the phrase-level features, such as co-occurrence counts or translation probabilities, and also propose a number of sentence-level context features. To calculate the sentence-level features for a given phrase-pair, we average the characteristics of all the sentences where it occurs.

We test the proposed algorithm experimentally, by constructing the bilingual lexicons for nine language pairs. The manually annotated samples of phrase pairs serve as the data for training supervised classifiers. The experiment shows that the use of the sentence-level features increases the classification accuracy, compared to a baseline which uses only phrase frequencies and translation probabilities. We compare the accuracy of different classifiers and evaluate the importance of different features.

The rest of the paper is organized as follows. In Section 2 we outline the related work. Section 3 describes our approach to the noise reduction in a bilingual lexicon and discusses the proposed features. We describe our experiments on training classifiers in Section 4. Section 5 concludes the paper.

2 Previous work

The methods of extracting a bilingual lexicon from parallel texts as a part of the alignment process are discussed in Brown (1993), Melamed (1996), Tufiş and Barbu (2001). Melamed (1996) proposes a method of noise reduction that allows

to re-estimate and filter out indirect word associations. However, he works with a carefully prepared Hansards parallel corpus and the noise comes only from the imperfections of statistical modeling.

Sahlgren (2004) proposes a co-occurrence-based approach, representing words as high-dimensional random index vectors. The vectors of translation equivalents are expected to have high correlation. Yet, he notes that low-frequency words do not produce reliable statistics for this method.

The methods of bilingual lexicon extraction from comparable texts (Rapp, 1995; Fung, 1998; Otero, 2007) also deal with the problem of noise reduction. However, the precision/recall ratio of a lexicon extracted from comparable corpus is generally lower. For the purpose of building a human-oriented dictionary, the parallel texts may provide the larger coverage and better quality of the translation equivalents.

The noise reduction task is addressed by some of the SMT phrase-table pruning techniques. The most straightforward approach is thresholding on the translation probability (Koehn et al., 2003). Moore (2004) proposes the log-likelihood ratio and Fisher's exact test to re-estimate word association strength. Johnson et al. (2007) applies Fisher's exact test to dramatically reduce the number of phrase pairs in the phrase-table. They get rid of phrases that appear as alignment artifacts or are unlikely to occur again. The implementation of their algorithm requires a special index of all parallel corpus in order to enable a quick look-up for a given phrase pair. Eck et al. (2007) assesses the phrase pairs based on the actual usage statistics when translating a large amount of text. Entropy-based criteria are proposed in Ling et al. (2012), Zens et al. (2012).

Automatically acquired bilingual lexicons are capable to reflect many word meanings and translation patterns, which are often not obvious even to the professional lexicographers (Sharoff, 2004). Their content can also be updated regularly to incorporate more parallel texts and capture the translations of new words and expressions. Thus, the methods allowing to improve the quality of automatic bilingual lexicons are of practical importance.

3 Noise detection features

We treat the noise recognition task as a binary classification problem. A set of nonlexical context features is designed to be sensitive to different types of noise in the parallel corpus. We expect that the combination of these features with the phrase-level features based on co-occurrence statistics can improve the accuracy of the classification and the overall quality of a bilingual lexicon.

3.1 Context feature extraction algorithm

The procedure of getting the context features is outlined in Algorithm 1. Unlike Johnson et al. (2007) we do not rely on any pre-constructed index of the parallel sentences, because it requires a lot of RAM on large corpora. Instead we re-run the phrase extraction algorithm of the Moses toolkit (Koehn et al., 2007) and update the context features at the moment when a phrase pair t is found.

Algorithm 1 Calculate context features for all lexicon entries

Require: Parallel corpus — C; {word-aligned sentences}
Require: Bilingual lexicon — D; {this is a phrase-table, derived from C and modified as descibed in 4.1}
Ensure: $V = \{\bar{v}(d) \colon d \in D\}$; {resulting features}
  for all $d \in D$ **do**
    $\bar{v}(d) \leftarrow 0$;
    $n(d) \leftarrow 0$;
  for all $s \in C$ **do**
    $T \leftarrow PhraseExtraction(s)$;{Moses function}
    for all $t \in T$ **do**
      if $t \in D$ **then**
        $\bar{v}(t) \leftarrow \bar{v}(t) + SentFeats(s)$; {Alg. 2}
        $n(t) \leftarrow n(t) + 1$;
  for all $d \in D$ **do**
    $\bar{v}(d) \leftarrow \bar{v}(d)/(1 + n(d))$; {average, +1 smoothing}
  return V

3.2 Sentence-level features

The phrase extraction algorithms do not preserve the information about the sentences in which a given phrase pair occurred, assuming that all the sentences are equally good. As a result, the

phrase-level statistics is insufficient in case of a noisy corpus.

The sentence-level features are designed to partly restore the information which is lost during the phrase extraction process. We try to estimate the general characteristics of the whole set of parallel sentences where a given phrase pair occurred. The proposed sentence-level features rely on the different sources of information, which are discussed in 3.2.1, 3.2.2 and 3.2.3. Table 2 provides illustrating examples of noisy phrase pairs and sample sentences.

3.2.1 Word-alignment annotation

We use the intersection of direct and reverse Giza++ (Och and Ney, 2004) alignments as a heuristic rule to find words reliably aligned to each other. The alignment information gives rise to several sentence-level features:

- $UnsafeAlign$ - percentage of words that are not symmetrically aligned to each other.
- $UnsafeJump$ - average distance between the translations of subsequent input words.
- $UnsafeDigAlign$ percentage of unequal digits among the symmetrically aligned words.

The $UnsafeAlign$ and $UnsafeJump$ values can vary in different sentences. However, their being too large on the whole set of sentences where a given phrase pair occurred possibly indicates some systematic noise.

The translations of digits are not included to the dictionary by themselves. But if a pair of digits is wrongly aligned, then its nearest context may also be aligned wrongly.

3.2.2 One-side morphological and syntactic annotation

The target side of our parallel sentences has been processed by a rule-based parser. The syntax gives rise to:

- $UnsafeStruct$ - percentage of words having no dependence on any other word in the parse tree.

The morphological annotation participates in:

- OOV - percentage of out-of-vocabulary words in the sentence.

The low parse tree connectivity may indicate that the sentence is ungrammatical or produced by a poor-quality machine translation system. Sentences containing many out-of-vocabulary words probably do not belong to the given language. We compute out-of-vocabulary words according to an external vocabulary, which is embedded in tagging and parsing tools. However, instead one can use a collection of unigrams filtered by some frequency threshold..

gratuit — internet access, $S_{lem} = 215$ Sample sentence: *Petit déjeuner continental de luxe gratuit* *Business center with free wireless Internet access* $UnsafeAlign = 0.387$
à — you, $S_{lem} = 586$ *La plainte à transmettre* *You should submit your complaint* $UnsafeJump = 1.75$
juin — May, $S_{lem} = 35$ *Membre depuis: 17 juin 2011* *Member since: 01 May 2012* $UnsafeAlignDig = 0.08$
le — Fr, $S_{lem} = 24$ *Edvaldo et le père Antenore* *Edvaldo and Fr Antenore* $OOV = 0.117$
Paris — England, $S_{lem} = 54$ *TERTIALIS (Paris, Paris)* *(England)* $Punct = 0.117$

Table 2: Examples of noisy French-English translations to which different sentence-level features may be sensitive. S_{lem} — is the number of sentences where a lemmatized phrase pair co-occurred. Sample sentences are provided.

3.2.3 Surface text

The surface word tokens can be used for:

- $Punct$ - percentage of non-word/punctuation tokens in the sentence.
- $Uniqueness$ - the percentage of unique unigrams in both source and target language sentences.

Sentences with lots of punctuation can be unnatural or contain enumeration. Large enumeration lists are often not exactly parallel and can be

aligned incorrectly, because punctuation tokens, like many commas, are easily mapped to each other. The low $Uniqueness$ possibly indicates that the sentences containing a given phrase pair are similar to each other. This can lead to overestimated translation probabilities.

Algorithm 2 Get features of one sentence pair ($SentFeats$)

Require: $sent_{src} = (w_1, ..., w_m)$;
Require: $sent_{dst} = (w_1, ..., w_n)$;
Require: Alignment matrix — $M_{m,n} : x \in \{0, 1\}$; {intersection of two Giza++ alignments}
Require: $oov = (x_1, ..., x_n), x \in \{0, 1\}$; $\{x_i = 1 \iff sent_{dst}[i]$ is out-of-vocabulary$\}$
Require: $pnt = (x_1, ..., x_n), x \in \{0, 1\}$; $\{x_i = 1 \iff sent_{dst}[i]$ is punctuation$\}$
Require: $nohead = (x_1, ..., x_n), x \in \{0, 1\}$; $\{x_i = 1 \iff sent_{dst}[i]$ is not dependent on any other word in the parse$\}$
Ensure: $\bar{v} = (v_1, ..., v_7)$; {features}
$\bar{v} \leftarrow 0$;
$v_2 \leftarrow \frac{1}{n} \sum_{x \in nohead} x$; $\{UnsafeStruct\}$

Let A be the set of pairs of indices of symmetrically aligned words, ordered by the source indices:
$A \leftarrow \{(i, j) \mid M(i, j) = 1\}$;

$v_3 \leftarrow 1 - \frac{|A|}{m+n}$; $\{UnsafeAlign\}$

for all $(i, j) \in A$ **do**
  if words with indices i, j are unequal digits **then**
    $v_4 \leftarrow v_4 + 1$;

$v_4 \leftarrow \frac{v_4}{|A|}$; $\{UnsafeAlignDig\}$

$v_5 \leftarrow \frac{1}{|A|} \sum_{(i,j) \in A} j_i - j_{i-1}$; $\{UnsafeJump\}$

$v_6 \leftarrow \frac{1}{n} \sum_{x \in oov} x$; $\{OOV\}$

$v_7 \leftarrow \frac{1}{n} \sum_{x \in pnt} x$; $\{Punct\}$

return $\bar{v}$

3.3 Phrase-level statistics

Multiple phrase-level features can be derived from the occurrence and co-occurrence counts, that are calculated during the phrase extraction procedure as described in Koehn et. al (2003).

- $C(f), C(e), C(e, f)$ — surface phrase occurrence counts.
- $C_{lem}(f)$, $C_{lem}(e)$, $C_{lem}(e, f)$ — same for lemmatized phrases.
- $S(e, f)$, $S_{lem}(e, f)$ — the number of sentences, in which the surface (or lemmatized) phrases co-occurred.
- $P(e|f)$, $P(f|e)$ — translation probabilities of surface phrases.
- $P_{lem}(e|f)$, $P_{lem}(f|e)$ — translation probabilities of lemmatized phrases.

Some of these features are highly correlated, and it is hard to tell in advance which subset leads to better performance.

4 Experiment

We conducted experiments on nine language pairs: German-English, German-Russian, French-English, French-Russian, Italian-English, Italian-Russian, Spanish-English, Spanish-Russian and English-Russian. The parallel corpora consisted of the sentence-aligned documents automatically collected from multilingual web-sites.

We implemented the procedure of bilingual lexicon construction and the algorithm calculating the sentence-level features (Section 3).

The annotated phrase pair samples, one for each language pair, provided positive and negative examples for training a supervised classifier. We compared the accuracy of several classifiers trained on different feature sets. The importance of different features was evaluated .

4.1 Bilingual lexicon creation

We used Giza++ for word alignment and Moses toolkit for phrase extraction procedure. The following automatic annotation had been provided. The source side of the parallel corpora had been processed by a part-of-speech tagger, and each word had been assigned a lemma based on its tag. The target side of the parallel corpora, which was always either English or Russian, was processed by a rule-based dependency parser, which also supplied morphological annotations and lemmas. In the case of English-Russian corpus, the source side had also been processed by the parser.

The extracted English phrases were restricted to at most 3 words, provided that they were connected in the dependency tree. The same restrictions were imposed on the Russian phrases. The extracted phrases for all other languages were restricted to single words to avoid the ungrammatical multiword expressions.

Each extracted phrase pair was assigned a lemmatized key consisting of lemmas of all words in it. The co-occurrence counts were summed over all phrase pairs sharing the same key, giving the aggregate count $C_{lem}(e, f)$. Then a single pair was chosen to serve as a best substitute for a lemmatized lexicon entry. The choice was made heuristically, based on the morphological attributes and co-occurrence counts.

As a preliminary lexicon cleanup we removed the phrase pairs which contained punctuation symbols or digits on either side. We also removed the pairs that co-occurred only once in the corpus. An example of differences between the size of original phrase table and the size of bilingual lexicon after lemmatization and preliminary cleanup is represented in Table 3.

	Millions of phrase pairs	
	fr-en	fr-ru
Initial 1-3 phrase-table	16.4	30.8
After lemmatization	7.9	6.4
After preliminary cleanup	1.6	0.8

Table 3: The number of phrase pairs on different stages of French-English and French-Russian dictionary creation. Phrase pairs in the initial phrase table are restricted to at most 1 source word and at most 3 target words.

4.2 Experimental data

For the experiment we selected random[1] translation equivalents from the nine translation lexicons, to which no further noise reduction had been applied. The resulting translation equivalents were assessed by human experts. The annotation task was to determine how well a phrase pair fits for a human-oriented translation dictionary. The annotators classified each translation according to the following gradation:

Class 0 — difficult to assess.

Class 1 — totally wrong or noisy (e.g. misspelled);

Class 2 — incorrect or incomplete translation;

Class 3 — not a mistake, but unnecessary translation;

Class 4 — good, but not vital;

Class 5 — vital translation (must be present in human-built dictionary);

The pairs annotated as 0 usually represented the translations of unfamiliar words, abbreviations and the like. Such phrases were excluded from training and testing. We didn't use "acceptable, but unnecessary" translation pairs either, because they do not influence the quality of the lexicon. We treated as negative the phrase pairs that were annotated as 1 or 2. Analogously, the positive examples had to belong to 4 or 5 class. The annotation statistics is given in Table 4.

Language	Size	%Negative	%Positive
it-ru	2340	56.6	28.7
it-en	2366	59.9	21.4
es-ru	2388	55.5	27.2
es-en	2384	69.0	24.0
de-ru	2397	50.3	37.6
de-en	2438	72.1	24.5
fr-ru	2461	44.5	31.2
fr-en	2325	57.0	24.4
en-ru	2346	27.8	33.2

Table 4: Statistics of the annotated data: the number of annotated phrase pairs, the percentage of negative and positive examples.

4.3 Training setting

The experiments were run with two different feature sets:

- Baseline — features based on co-occurrence counts.
- Full — baseline and sentence-level features.

We had to choose a subset of co-occurrence-based features experimentally (see, Section 3.3). The best subset for our data consisted of three features: $log(S_{lem})$, $log(P(e|f))$, $log(P(f|e))$. In the full feature set we combined the baseline features and the sentence-level features calculated as described in Algorithm 2.

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

We considered three metrics related to the improvement of the lexicon quality:

- Err — the percentage of prediction errors;
- Err-1 — the percentage of class 1 examples which were classified as positive.
- F1 — the harmonic mean of precision and recall w.r.t. the positive and negative examples;

We used the standard packages of the R programming language, to train and tune different classifiers: random forest (RF), support vector machines (SVM), logistic regression (GLM), Naive Bayes classifier, neural networks, k-Nearest Neighbors and some of the combinations of these methods with SVD. To assess the predictive accuracy we used repeated random sub-sampling validation. In each of 40 iterations, a 10% test set was randomly chosen from the dataset, the model was trained on the rest of the data, and then tested. The resulting accuracy was averaged over the iterations.

Classifier	Full feature set		Base feature set	
	%Err	%Err-1	%Err	%Err-1
RF	19.80	8.31	24.00	14.62
SVM	19.63	9.36	23.49	12.91
GLM	22.74	6.35	25.23	7.30

Table 5: Percentage of prediction errors of different classifiers, averaged over the nine language pairs.

The results of RF, SVM and GLM are reported in Table 5. Though the composition of different classifiers could perform slightly better, it would require an individual tuning for each language pair. For clearness, we use a single classifier (RF) for the rest of the experiments.

The experiment showed that training on the full feature set reduced the total amount of prediction errors by 17.5%, compared to the baseline setting. The number of false positives among the class 1 examples reduced by 43%. It is also important that better results were obtained on each of the nine language pairs, not only on average. In Table 6 the baseline results are shown in brackets and one can see that F1 diminishes in the baseline setting, while the percentage of errors goes up. The classification accuracy depends on the size of the training set (Table 7).

Lang	%Err	%Err-1	F1
de-en	18.0 (+3.6)	4.0 (+5.2)	.562 (-.050)
de-ru	25.7 (+4.0)	13.5 (+6.7)	.672 (-.040)
es-en	16.4 (+3.8)	3.2 (+4.0)	.610 (-.059)
es-ru	20.6 (+4.7)	8.3 (+6.0)	.643 (-.064)
fr-en	20.5 (+1.5)	6.0 (+5.8)	.603 (-.031)
fr-ru	21.4 (+6.1)	15.5 (+10.8)	.704 (-.070)
it-en	15.2 (+3.3)	3.5 (+2.9)	.663 (-.059)
it-ru	19.6 (+5.5)	9.4 (+6.7)	.670 (-.071)
en-ru	20.8 (+5.6)	11.5 (+8.8)	.797 (-.048)

Table 6: Classification quality of the classifier trained on all features, compared to the baseline trained only on phrase-level features. The relative change of the baseline values is given in brackets.

Examples	1700	680	272	108	43
Accuracy	.803	.794	.780	.757	.709

Table 7: Classification accuracy w.r.t different size of training set averaged over eight language pairs.

We measured the impact of different features, as described in Breiman (2001), with the help of the standard function of the R library "randomForest" (Table 8). The three baseline features were ranked as most important, followed by UnsafeAlign, OOV, UnsafeJump and others.

Feature	Importance
$log(S_{lem})$	35.679
$log(P(e\|f))$	33.9729
$log(P(f\|e))$	28.8637
$UnsafeAlign$	24.3705
OOV	22.8306
$UnsafeJump$	20.1108
$Punct$	15.4501
$UnsafeStruct$	15.1157
$Uniqueness$	13.5049
$UnsafeDigAlign$	12.915

Table 8: Feature importance measured by the mean decrease of classification accuracy (Breiman, 2001). The value is averaged over the nine language pairs.

We explored the dependence of the prediction accuracy on the co-occurrence frequency of a phrase pair for the classifiers trained on the full feature set and on the baseline feature set. The results for German-English and French-English lan-

guage pairs are shown in Figure 2. The accuracy function was smoothed with cubic smoothing spline. The differences in the distribution of classification errors between language pairs suggest that the nature of the noise can vary for different corpora. The general U shape of the curves in Figure 2 is partly due to the fact that there are many true negatives in the low-frequency area, and many true positives in the high-frequency area.

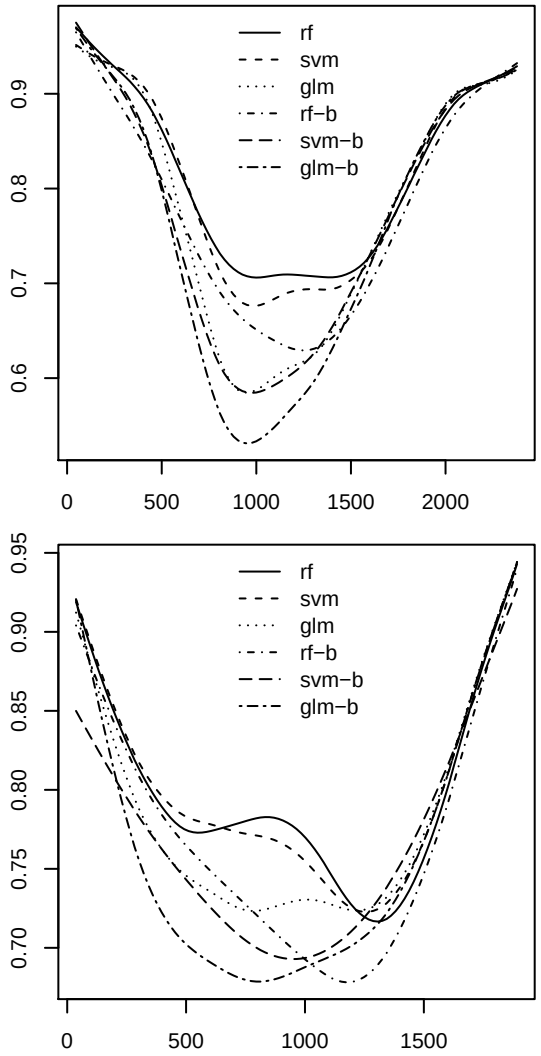

Figure 2: Prediction accuracy of different classifiers w.r.t. the phrase pairs sorted by the ascending co-occurrence count. The upper plot relates to the German-English pair, the bottom relates to French-English pair. The labels rf, svm, glm refer to the classifiers trained on the full feature set; rf-b, svm-b, glm-b refer to the baseline setting.

Table 9 reports the top English translations of the French word "connexion" before the noise reduction and shows which variants were recognized as positive and negative by the RF classifier.

English	$C(e,f)$	$p(f\|e)$	$p(e\|f)$	RF
connection	58018	0.689	0.374	+
wireless	32630	0.450	0.211	-
free	31775	0.113	0.205	-
wifi	16272	0.382	0.105	-
login	4910	0.443	0.032	+
connectivity	394	0.055	0.003	+
logon	290	0.185	0.002	+
access	276	0.001	0.002	-
link	148	0.001	0.001	-

Table 9: English translations of the French word "connexion". $C(e,f)$ is the co-occurrence count, $p(f|e)$, $p(e|f)$ are the translation probabilities of lemmatized pairs. The last column shows the classification result.

5 Conclusion

The main contributions of this paper are the following. We address the problem of noise reduction in automatic construction of human-oriented translation dictionary. We introduce an approach to increase the precision of automatically acquired bilingual lexicon, which allows to mitigate the negative impact of a noisy corpus. Our noise reduction method relies on the supervised learning on a small set of annotated translation pairs. In addition to the phrase-level statistics, such as co-occurrence counts and translation probabilities, we propose a set of non-lexical context features based on the analysis of sentences in which a phrase pair occurred. The experiment demonstrates a substantial improvement in the accuracy of the detection of noisy translations, compared to a baseline which uses only phrase-level statistics.

We have shown that the proposed noise detection method is applicable to various language pairs. The alignment-based features can be easily obtained for any parallel corpus, even if other tools do not exist. We hope that our noise detection approach can also be adapted for SMT phrase-tables, if the initial parallel sentences are still available.

References

B. T. Sue Atkins. 1994. *A corpus-based dictionary*. In Oxford-Hachette French Dictionary, Introduction xix-xxxii. Oxford: Oxford University Press.

Leo Breiman. 2001. *Random Forests*. Machine Learning 45 5-32.

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.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2007. *Translation model pruning via usage statistics for statistical machine translation*. In Human Language Technologies 2007: The Conference of the NAACL; Companion Volume, Short Papers, pages 21–24, Rochester, New York, April. Association for Computational Linguistics

Pascale Fung. 1998. *A Statistical View on Bilingual Lexicon Extraction from Parallel Corpora to Non-parallel Corpora*. Parallel Text Processing: Alignment and Use of Translation Corpora. Kluwer Academic Publishers

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.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn 2007. *Improving translation quality by discarding most of the phrasetable*. In Proceedings of EMNLP-CoNLL, ACL, Prague, Czech Republic, pages 967-975.

Philipp Koehn, Franz Josef Och, and Daniel Marcu 2003. *Statistical phrase-based translation*. In Proceedings of HLT-NAACL 2003, pages 127–133.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej 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

Akira Kumano and Hideki Hirakawa. 1994. *Building An MT Dictionary From Parallel Texts Based On Linguistic And Statistical Information*. COLING 1994: 76-81

Wang Ling, João Graça, Isabel Trancoso and Alan Black 2012. *Entropy-based Pruning for Phrase-based Machine Translation*. In Proceedings of EMNLP-CoNLL, Association for Computational Linguistics, Jeju Island, Korea, pp. 972-983

C. J. A. McEwan, I. Ounis, and I. Ruthven. 2002. *Building bilingual dictionaries from parallel web documents*. In Proceedings of the 24th BCS-IRSG European Colloquium on IR Research, pp. 303-323. Springer-Verlag.

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

I. Dan Melamed. 2000. *Models of Translational Equivalence among Words*. Computational Linguistics 26(2), 221-249, June.

Robert C. Moore. 2004. *On Log-Likelihood-Ratios and the Significance of Rare Events*. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain.

Franz Josef Och and Hermann Ney. 2000. *Improved Statistical Alignment Models*. Proceedings of the 38th Annual Meeting of the ACL, pp. 440-447, Hongkong, China.

Franz Josef Och and Hermann Ney. 2004. *The Alignment Template Approach to Statistical Machine Translation*. Computational Linguistics, vol. 30 (2004), pp. 417-449.

Pablo Gamallo Otero. 2007. *Learning bilingual lexicons from comparable English and Spanish corpora*. Proceedings of MT Summit XI, pages 191–198.

Reinhard Rapp. 1995. *Identifying word translations in non-parallel texts*. In Proceedings of the ACL 33, 320-322.

Resnik, Philip and Noah A. Smith. 2003. *The web as a parallel corpus.*. Computational Linguistics, 29, pp.349–380

Magnus Sahlgren. 2004. *Automatic Bilingual Lexicon Acquisition Using Random Indexing*. Journal of Natural Language Engineering, Special Issue on Parallel Texts, 11.

Serge Sharoff. 2004. *Harnessing the lawless: using comparable corpora to find translation equivalents*. Journal of Applied Linguistics 1(3), 333-350.

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.

Dan Tufiş 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

Velupillai, Sumithra, Martin Hassel, and Hercules Dalianis. 2008. *"Automatic Dictionary Construction and Identification of Parallel Text Pairs*. In Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS).

Richard Zens, Daisy Stanton and Peng Xu. 2012. *A Systematic Comparison of Phrase Table Pruning Techniques*. In Proceedings of EMNLP-CoNLL, ACL, Jeju Island, Korea, pp. 972-983.