Unsupervised training of maximum-entropy models for lexical selection in rule-based machine translation

Francis M. Tyers
HSL-fakultehta,
UiT Norgga árktalaš universitehta,
N-9018 Romsa

Felipe Sánchez-Martínez
Dept. Lleng. i Sist. Inform.,
Universitat d'Alacant,
E-03071 Alacant

Mikel L. Forcada
Dept. Lleng. i Sist. Inform.,
Universitat d'Alacant,
E-03071 Alacant

Abstract

This article presents a method of training maximum-entropy models to perform lexical selection in a rule-based machine translation system. The training method described is unsupervised; that is, it does not require any annotated corpus. The method uses source-language monolingual corpora, the machine translation (MT) system in which the models are integrated, and a statistical target-language model. Using the MT system, the sentences in the source-language corpus are translated in all possible ways according to the different translation equivalents in the bilingual dictionary of the system. These translations are then scored on the target-language model and the scores are normalised to provide fractional counts for training source-language maximum-entropy lexical-selection models. We show that these models can perform equally well, or better, than using the target-language model directly for lexical selection, at a substantially reduced computational cost.

1 Introduction

Corpus-based machine translation (MT) has been the primary research direction in the field of MT in recent years. However, rule-based MT (RBMT) systems are still being developed, and there are many successful commercial and non-commercial systems. One reason for the continued development of RBMT systems is that in order to be successful, corpus-based MT requires parallel corpora in the order of tens of millions of words. Although for some language pairs these exist, they only exist for a fraction of the world's languages.

An RBMT system typically consists of an analysis component,[1] a transfer component and a generation component. As part of the transfer component it is necessary to make choices regarding words in the source language (SL) which may have more than one translation in the target language (TL).

Lexical selection is the task of choosing, for a given SL word, the most adequate translation in the TL among a known set of alternatives. The task is related to the task of word-sense disambiguation (Ide and Véronis, 1998). However, it is different to word-sense disambiguation in that lexical selection is a bilingual problem, not a monolingual problem: its aim is to find the most adequate translation, not the most adequate sense. Thus, it is not necessary to choose among a series of fine-grained senses if all these senses result in the same final translation; however, it may sometimes be necessary to choose a different translation for the same sense, for example in a collocation.

1.1 Prior work

Dagan and Itai (1994) used the term *word sense disambiguation* to refer to what is actually lexical selection in MT; they used a parser to identify syntactic relations such as subject–object or subject–verb. After generating all the possible translations for a given input sentence using an ambiguous bilingual dictionary, they extract the syntactic tuples from the TL and count the frequency in a previously-trained TL model of tuples. They use maximum-likelihood estimation to calculate the probability that a given

[1] Such as a morphological or syntactic analyser.

TL tuple is the translation of a given SL tuple, with an automatically determined confidence threshold.

Later, Berger et al. (1996) illustrated the use of maximum-entropy classifiers on the specific problem of lexical selection in IBM-style word-based statistical MT. Other authors (Melero et al., 2007) have used TL models to rank the translations resulting from all possible combinations of lexical selections. Nowadays, in state-of-the-art phrase-based statistical MT (Koehn, 2010), lexical selection is taken care of by a combination of the translation model and the language model. The translation model provides probabilities of translation between words or word sequences (often referred to as *phrases*) in the source and target language. The TL model provides probabilities of word sequences in the TL. Mareček et al. (2010) trained a maximum-entropy lexical selector for their dependency-grammar-based transfer system TectoMT using a bilingual corpus. More recently, Tyers et al. (2012) presented a method of lexical selection for RBMT based on rules which select or remove translations in fixed-length contexts, along with a training method for learning the rules from a word-aligned parallel corpus.[2]

2 Method

Lexical selection in this paper considers for each word a simple SL context made up of neighbouring lemma+part-of-speech combinations. Contexts considered include up to two words to the left and up to two words to the right of the word to be translated.

Let the probability of a word t being the translation of a word s in a SL context c be $p_s(t|c)$. In principle, this value could be estimated directly from the available corpora for every combination of (s, t, c). This would however present two questions: (1) how should the relevant contexts be chosen? and (2) what should be done when (s, t, c) is not found in the corpus? A maximum-entropy model answers both of these questions. It allows the contexts that we consider to be linguistically interesting to be defined *a priori* and then integrate these seamlessly into a probabilistic model (Manning and Schütze, 1999). In answer to the second question, a maximum-entropy model maximises the entropy subject to match the expected counts of the designed features with those found in the training data. That is, if there is no information in the training data, then it assumes that all outcomes —that is, all possible translations— are equally likely. As previously mentioned, the principle of maximum entropy has been applied to the problem of lexical selection before; in particular, Berger et al. (1996) cast the problem of lexical selection in statistical MT as a classification problem. They learn a separate maximum-entropy classifier for each SL word form, using SL context to distinguish between possible translations. These classifiers are then incorporated into the translation model of their word-based statistical MT system. In their approach, a classifier consists of a set of binary feature functions and corresponding weights for each feature. In both Berger et al. (1996) and our method, features are defined in the form $h_k^s(t, c)$,[3] where t is a translation, and c is a SL context. One difference is that Berger et al. (1996) take s, t and c to be based on word forms, whereas in our method they are based on lemma forms. An example would be the following feature where the Spanish word *pez* ('fish' as a living animal) is seen as the translation of *arrain* ('fish') in the context *arrain handi* 'big fish' and would therefore be defined as:

$$h^{\text{arrain}}_{+\text{handi}}(t,c) = \begin{cases} 1 & \text{if } \begin{cases} t = \textit{pez} \\ \text{and} \\ \textit{handi} \text{ follows } \textit{arrain} \end{cases} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This feature considers a context of zero words to the left of the problem word and one word (+ *handi*) to the right of it.

As a result of training, each of the n_F features $h_k^s(t, c)$ in the classifier is assigned a weight λ_k^s. Combining these weights of active features as in equation (2) yields the probability of a translation t for word s in context c.

$$p_s(t|c) = \frac{1}{Z^s(c)} \exp \sum_{k=1}^{n_F} \lambda_k^s h_k^s(t, c) \tag{2}$$

In this equation, $Z^s(c)$ is a normalising constant. Thus, the most probable translation $t^\star$ can be found using

$$t^\star = \arg\max_{t \in T_s(s)} p_s(t|c) = \arg\max_{t \in T_s(s)} \sum_{k=1}^{n_F} \lambda_k^s h_k^s(t, c), \tag{3}$$

[2] The work by Ravi and Knight (2011) and Nuhn and Ney (2014), who decipher word-ciphered text using monolingual corpora only may be seen as a generalised version of the problem of lexical selection without parallel corpora.

[3] We follow the notation of Berger et al. (1996)

$$S \rightarrow \boxed{\textit{pre-lexsel}} \rightarrow (\{g_i\}_{i=1}^{i=|G|}, S) \rightarrow \boxed{\textbf{\textit{lexsel}}} \rightarrow (g^\star, S) \rightarrow \boxed{\textit{post-lexsel}} \rightarrow \tau(g^\star, S)$$

Figure 1: A schema of the lexical selection process: source sentence S has $|G|$ lexical selection paths g_i: *lexsel* selects one of them $g^\star$, which is used to generate translation $\tau(g^\star, S)$.

where $T_s(s)$ is the set of possible translations for SL word s.

The approaches by Berger et al. (1996) and by Mareček et al. (2010) cited above both take advantage of a parallel corpus to collect counts of contexts and translations in order to train maximum-entropy models. However, parallel corpora are not available for the majority of the world's written languages. In this section we describe an unsupervised method to learn the models using only monolingual corpora and the components from the RBMT system in which they are used.

The input to our method consists of a collection of samples, $\mathcal{G} = (S, G)$, where $S = (s_1, s_2, \ldots s_{|S|})$ is a sequence of SL words, and $G = \{g_1, g_2, \ldots g_{|G|}\}$ is a set of possible *lexical-selection paths*. A lexical-selection path $g = (t_1, t_2, \ldots, t_{|S|})$ is a sequence of lexical-selection choices of those SL words, where t_i is an element of $T_s(s_i)$, the set of possible translations of s_i.[4] This is produced in the first stages of RBMT, just after morphological analysis, part-of-speech tagging, and bilingual dictionary lookup, and before any structural transfer takes place (we will call this *pre-lexsel*). In our model, it is after these first stages that lexical selection (*lexsel*) occurs. After lexical selection, structural transfer and generation take place; a function $\tau(g_i, S)$ represents the result of these last stages, which we will call *post-lexsel*, and returns a finished translation of a specific lexical-selection path g_i of sentence S. Figure 1 shows this process schematically.

As our method is unsupervised, and therefore the occurrences of specific lexical selection events (s, t, c) cannot be counted, a TL model $P_{\mathrm{TL}}(\cdot)$ is used to compute a value for the fractional count for disambiguation path g_i, $p(g_i|S)$ after suitable normalisation:

$$p(g_i|S) = \frac{P_{\mathrm{TL}}(\tau(g_i, S))}{\sum_{g_i \in G} P_{\mathrm{TL}}(\tau(g_i, S))} \qquad (4)$$

The maximum-entropy model is trained instead using the fractional count $p(g_i|S)$ for the events (s, t, c) found in g_i, that is, when in g_i the translation for s in context c is t. That is, as if event (s, t, c) had been seen a fractional number $p(g_i|S)$ of times. We prune (s, t, c) occurring less than a certain number of times in the corpus, using a development corpus to guide pruning (see section 4). The method used here for lexical selection is analogous to the method used by Sánchez-Martínez et al. (2008) to train a hidden-Markov-model-based part-of-speech tagger in a RBMT system.

3 Experimental setting

This section describes the training and evaluation settings used in the remainder of this paper. The primary motivation behind the evaluation is that it should be automatic, meaningful, and be performed over a test set which is large enough to be representative. It should evaluate both performance on the specific subtask of lexical selection, and on the whole translation task. Evaluating lexical-selection performance is an *intrinsic* module-based evaluation. It measures how well the lexical selection module disambiguates the lexical-transfer output as compared to a gold-standard corpus. The lexical transfer output is the result of looking up the translations of the SL *lexical forms* — lemmas and tags — in the bilingual dictionary.

The whole translation task evaluation is an *extrinsic* evaluation, which tests how the system improves as regards final translation quality in a real system.

The lexical-selection module should be as language-independent as possible. To that end, the language pairs tested show a wide variety of linguistic phenomena. It is also important that the methodology be as applicable to lesser-resourced and marginalised languages as to major languages.

This section begins with a short description of the Apertium platform (Forcada et al., 2011). This is followed by an overview of each of the language pairs chosen for the evaluation. The corpora to be used for training and evaluation will subsequently be described, along with the method used for annotating them. This is followed by a description of the performance measures to be used in the evaluation, and the reference results using these metrics for

[4] We deal only with single-word translations in this paper.

each of the language pairs.

3.1 Apertium

Apertium is a free/open-source RBMT platform, it comprises an engine, a toolbox and data to build RBMT systems. Translation is implemented as a pipeline consisting of the following modules: morphological analysis, morphological disambiguation, lexical transfer, lexical selection, structural transfer and morphological generation.

3.2 Language pairs

Evaluation will be performed using four Apertium (Forcada et al., 2011) language pairs. These pairs have been selected as they include languages with different morphological complexity, and different amounts of resources available — although for all pairs there is a parallel corpus available for evaluation (see Section 3.3).[5]

Breton–French (Tyers, 2010): Bilingual dictionaries were not built with polysemy in mind from the outset, but some entries were added later to start work on lexical selection.[6]

Macedonian–English: The Macedonian–English pair in Apertium was created specifically for the purposes of running lexical-selection experiments. The lexical resources for the pair were tuned to the SETimes parallel corpus (Tyers and Alperen, 2010). The most probable entry from automatic word alignment of this corpus using GIZA++ (Och and Ney, 2003) was checked to ensure that it was an adequate translation, and if so marked as the default.[7] As a result of attempting to include all possible translations, the average number of translations per word is much higher than in other pairs.[8]

Basque–Spanish (Ginestí-Rosell et al., 2009): alternative translations were included in the bilingual dictionary.[9]

English–Spanish: The English–Spanish pair was developed from a combination of the English–Catalan and Spanish–Catalan pairs, and contains a number of entries in the bilingual dictionary with more than one translation.[10]

3.3 Performance measures

This section describes the measures that will be used to evaluate the performance of the lexical selection method proposed here: a (intrinsic) *lexical selection performance* measure and an (extrinsic) *machine translation performance* measure.

3.3.1 Lexical-selection performance

This is an intrinsic module-based evaluation of the performance of the lexical-selection module. It measures how well the lexical-selection module disambiguates the output of the lexical-transfer module as compared to a gold-standard corpus. For this task, we define a metric, the lexical-selection error rate (LER), that focuses on the problem of lexical selection by restricting the evaluation to this feature; other features of the MT system, such as the transfer rules and morphological generation, are not taken into account.

The lexical-selection error rate is the fraction of times the given system chooses a translation for a word which is not the one found in an annotated reference. The process uses a SL sentence, $S = (s_1, s_2, \ldots, s_{|S|})$ and three functions. The first function, $T_s(s_i)$, returns all possible translations of s_i according to the bilingual dictionary. The second function, $T_t(s_i)$, returns the translations of s_i selected by the lexical-selection module: $T_t(s_i) \subseteq T_s(s_i)$; and usually $|T_t(s_i)| = 1$. If the lexical-selection module returns more than one translation, the first translation is selected. The function $T_r(s_i)$ returns the set of reference translations which are acceptable for s_i in sentence S.[11] For a single sentence, we define the lexical selection error rate (LER) of that sentence as

$$\mathrm{LER} = \frac{\sum_{i=1}^{|S|} \mathrm{amb}(s_i)\ \mathrm{diff}(T_r(s_i), T_t(s_i))}{\sum_{i=1}^{|S|} \mathrm{amb}(s_i)}, \quad (5)$$

where

$$\mathrm{amb}(s_i) = \begin{cases} 1 & \text{if } |T_s(s_i)| > 1 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

[5] The Apertium revision (version) used is given in footnotes.

[6] Revision 41375; https://svn.code.sf.net/p/apertium/svn/trunk/apertium-br-fr

[7] Bilingual dictionaries in Apertium (Forcada et al., 2011) may contain several translations for a given word. Dictionary writers may mark as *linguistic default* the most general or most frequent translation among the set of possible translations.

[8] Revision 41476; https://svn.code.sf.net/p/apertium/svn/trunk/apertium-mk-en

[9] Revision 44846; https://svn.code.sf.net/p/apertium/svn/trunk/apertium-eu-es

[10] Revision 41387; https://svn.code.sf.net/p/apertium/svn/trunk/apertium-en-es

[11] Depending on how the reference is built, the set returned by $T_r(s_i)$ may not include all possible acceptable translations.

L'estiu és una estació llarga						
S	*el*	*estiu*	*ser*	*un*	*estació*	*llarg*
$T_s(s_i)$	{the}	{summer}	{be}	{a}	{station, season}	{long, lengthy}
$T_r(s_i)$	{the}	{summer}	{be}	{a}	{season}	{long}
$T_t(s_i)$	{the}	{summer}	{be}	{a}	{station}	{long}
amb(s_i)	0	0	0	0	1	1
diff$(T_r(s_i), T_t(s_i))$	0	0	0	0	1	0

Figure 2: An example input sentence in Catalan and the three sets of English translations used for calculating the lexical-selection error rate. The source sentence $S = (s_1, s_2, \ldots, s_{|S|})$ has two ambiguous words, *estació* and *llarg* (amb$(s_i) = 1$, eq. (6)). There is one difference (diff$(T_r(s_i), T_t(s_i)) = 1$, eq. (7)) between the reference set $T_r(s_i)$ and the test set $T_t(s_i)$ of translations; thus, the error rate for this sentence is 50%.

tests if a word is ambiguous, and the function

$$\text{diff}(T_r(s_i), T_t(s_i)) = \begin{cases} 1 & \text{if } T_r(s_i) \cap T_t(s_i) = \emptyset \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

states that there is a difference if the intersection between the set of reference translations $T_r(s_i)$ and the set of translations from the lexical selection module $T_t(s_i)$ is empty. Recall that, although $T_t(s_i)$ returns a set, this set will be a singleton, as when the lexical-selection module returns more than one translation, Apertium will select the default one if marked or the first one of not.[12]

The table in Figure 2 gives an overview of the inputs.In the description it is assumed that the reference translation has been annotated by hand. However, hand annotation is a time-consuming process, and was not possible. A description of how the reference was built is given in Section 3.4.

3.3.2 Machine translation performance

This is an extrinsic evaluation, which ideally would test how much the system improves as regards an approximate measurement of final translation quality in a real system. For this task, we use the widely-used BLEU metric (Papineni et al., 2002). This is not ideal for evaluating the task of a lexical selection module as the performance of the module will depend greatly on (a) the coverage of the bilingual dictionaries of the RBMT system in question, and (b) the number of reference translations. It is also worth noting that successful lexical selections may not lead to successful translations due to inadequate transfer of morphological features. The BLEU metric is included only as it is commonly used to evaluate MT systems.

[12]In practice this does not happen as each ambiguous word has a *default* translation.

3.3.3 Confidence intervals

Confidence intervals for both metrics will be calculated through *bootstrap resampling* (Efron and Tibshirani, 1994) as described by Koehn (2004). In all cases, bootstrap resampling will be carried out for 1,000 iterations. Where the $p = 0.05$ confidence intervals overlap, we will also perform paired bootstrap resampling (Koehn, 2004).

3.4 Corpora

For creating the test corpora, providing a SL corpus for training, and a TL corpus for scoring, we used four parallel corpora:

- **Ofis ar Brezhoneg** (`OAB`): This parallel corpus of Breton and French has been collected specifically for lexical-selection experiments from translations produced by *Ofis ar Brezhoneg* 'The Office of the Breton language'. The corpus has recently been made available online through OPUS.[13]
- **South-East European Times** (`SETimes`): Described in Tyers and Alperen (2010), this corpus is a multilingual corpus of the Balkan languages (and English) in the news domain. The Macedonian and English part will be used.
- **Open Data Euskadi** (`OpenData`): This is a Basque–Spanish parallel corpus made from the translation memories of the *Herri Arduralaritzaren Euskal Erakundea* 'Basque Institute of Public Administration'.[14]
- **European Parliament Proceedings** (`EuroParl`): Described by Koehn (2005), this is a multilingual corpus of the European Union official languages. We are using the English–Spanish data from version 7.[15]

[13]`http://opus.lingfil.uu.se`
[14]`http://tinyurl.com/eu-es-tm`
[15]`http://www.statmt.org/europarl/`

There are a number of approaches to creating evaluation corpora for lexical selection in the literature. Vickrey et al. (2005) use a parallel corpus to make annotated test and training sets for experiments in lexical selection applied to a simplified translation problem in statistical MT. They use word alignments from GIZA++ (Och and Ney, 2003) to annotate SL words with their translations from the reference translation in the parallel corpus. One disadvantage of this method is that only one translation is annotated per SL word, meaning that accuracies may be lower because of missing translations — this happens when the system chooses a translation which is adequate, but is not found in the reference translation. A second disadvantage is that the word alignments may not be 100% reliable, which decreases the accuracy of the annotated corpus. An alternative method is described by Zinovjeva (2000), who manually tags ambiguous words in English sentences with their translation in Swedish.

Ideally we would have had a hand-annotated evaluation corpus, as described by Zinovjeva (2000), but as this did not exist, we decided to automatically annotate a test set using a process similar to that described by Vickrey et al. (2005).

The annotation process proceeds as follows: First we word-align the corpus to extract a set of word alignments, which are correspondences between words in sentences in the source side of the parallel corpus and those in the target side. Any aligner may be used, but in this paper we use GIZA++ (Och and Ney, 2003).[16] We then use these alignments along with the bilingual dictionary of the MT system in question to extract only those sentences where: (a) there is at least one ambiguous word; (b) that ambiguous word is aligned to a single word in the TL; and (c) the word it is aligned to in the TL is found in the bilingual dictionary of the MT system. Sentences where there are no ambiguous words (approximately 90%, see Table 1) are discarded. The source side of the extracted sentence is then passed through the lexical transfer module, which returns all the possible translations, and for each ambiguous word, the translation is selected which is found aligned in the reference.

After this process, we selected 1,000 sentence pairs at random for testing (`test`), 1,000 for development (`dev`)[17] and left the remainder for training. Table 1 gives statistics about the size of the input corpora, and how many sentences were left after processing for testing, training and development. Table 2 gives information about the test corpora.

Pair	SL	TL	Amb.	% amb.
br-fr	13,854	13,878	1,163	8.39
mk-en	13,441	14,228	3,872	28.80
eu-es	7,967	11,476	1,360	17.07
en-es	19,882	20,944	1,469	7.38

Table 2: Statistics about the test corpora. The columns **SL** and **TL** give the number of tokens in the source and target languages respectively. The columns **amb. words** and **% ambig** gives the number of word with more than one translation and the percentage of SL words which have more than one translation respectively.

3.5 Reference systems

We compare our method to the following reference (or baseline) systems:

- **Linguist-chosen defaults**. A bilingual dictionary in an Apertium language pair contains correspondences between lexical forms. The dictionaries allow many lexical forms to translate to one lexical form. But a single lexical form may not have more than one translation without further processing. If there are many possible translations of a lexical form, then one must be marked as the *default* translation.
- **Oracle**. The results for the oracle system are those achieved by passing the automatically annotated reference translation through the rest of the modules of the MT system. This is included to show the upper bound for the performance of the lexical-selection module.
- **Target language model** (TLM). One method of lexical selection is to use the existing MT system to generate all the possible translations for an input sentence, and then score these translations *on-line* on a model of the TL. The highest scoring sentence is then output. This is the method used by Melero et al. (2007).

4 Results

As we are working with binary features, we use the implementation of generalised iterative scaling

[16] The exact configuration of GIZA++ used is equivalent to running the MOSES toolkit (Koehn et al., 2007) in default configuration up to step three of training.

[17] The development corpus was used for checking the value for frequency pruning of features.

Pair	Lines	Extract.	train	dev	test	No. amb	Av. amb
br-fr	57,305	4,668	2,668	1,000	1,000	603	3.06
mk-en	190,493	19,747	17,747	1,000	1,000	13,134	3.06
eu-es	765,115	87,907	85,907	1,000	1,000	1,806	3.11
en-es	1,467,708	312,162	310,162	1,000	1,000	2,082	2.28

Table 1: Statistics about the source corpora. The column **no. amb** gives the number of unique tokens with more than one possible translation. The column **av. amb** gives the average number of translations per ambiguous word. This is calculated by looking up each word in the corpus in the bilingual dictionary of the MT system and dividing the total number of translation by the number of words. Both **av. amb** and **no. amb** are calculated over the whole corpus.

Pair	Pruned	# features
br-fr	< 5	5,277
mk-en	< 7	205,494
eu-es	< 7	196,024
en-es	< 7	195,605

Table 3: Features in each rule set and pruning frequency.

available in the YASMET[18] to calculate the feature weights. After learning the feature sets and weights, we compute the evaluation measures described in Section 3.3. There is an option to remove events (s, t, c) which occur less than a certain number of times in the training corpus. This is referred to as the feature pruning frequency threshold — features occurring less than the threshold are discarded. The value was set experimentally. Values of between two and seven were tested, and the ones which provided the best improvement on the development corpus were selected; they happen to come close to the rule-of-thumb value of five that Manning and Schütze (1999, p. 596) found to be effective. Table 3 shows the number of features that have eventually been used for each language pair.

Evaluation results are presented in table 4, which compares the results of the new approach with respect to the *default behaviour* (the linguist-chosen defaults), with respect to the *oracle* (which represents the upper bound to performance), and with respect to the results obtained by using the TL model online, for each of the language pairs in Apertium with respect to our two evaluation metrics. Note that the high error rate for the Breton–French pair may be as a result of having the linguistic defaults tuned to a different domain than that of the corpus.

Significant improvements with respect to the results obtained using the TL model online are apparent with the Breton–French —the pair with the least data— and the English–Spanish language pairs. In the remaining cases, the maximum-entropy method comes close to the TL model performance in terms of similar or better BLEU and LER scores, at a much smaller computational cost.

Improvements with respect to the TL model performance are likely due to the effective use that the maximum-entropy model makes of information about the relevant SL contexts and their translations, through the weighting of features representing those SL contexts across the whole corpus.

5 Conclusions

This paper has presented a method to perform lexical selection in RBMT, and one that can be trained in an unsupervised way, that is, without the need for an annotated corpus, (in this case a word-aligned bilingual corpus): one just needs a SL corpus, a statistical TL model, and the RBMT system itself. The input to the method is simply the part-of-speech tagged source text in which each word is annotated with all the translations provided by the bilingual dictionary in the system: this makes it applicable to almost any RBMT system. The system uses a maximum-entropy formalism for lexical selection, as Berger et al. (1996) and Mareček et al. (2010), but instead of counting actual lexical selection events in an annotated corpus, it counts fractional occurrences of these events as estimated by a TL model. The method is evaluated both intrinsically (just looking at the actual lexical selection events) and extrinsically (measuring the quality of MT). Results on four language pairs using the Apertium (Forcada et al., 2011) MT system show that the method obtains similar or better results than those expensively obtained by scoring an exponential number of lexical selections for each sentence using the TL model online.

[18] http://www-i6.informatik.rwth-aachen.de/web/Software/YASMET.html; the compilable version we used is available as part of the Apertium lex-tools package, http://downloads.sourceforge.net/project/apertium/apertium-lex-tools/apertium-lex-tools-0.1.0.tar.gz.

Pair	Metric	System			
		Ling	TLM	MaxEnt	Oracle
br-fr	LER (%)	[54.8, 60.7]	[44.2, 50.5]	**[40.8, 46.9]**	[0.0, 0.0]
	BLEU (%)	[14.5, 16.4]	[15.4, 17.3]	[14.8, 16.6]	[16.7, 18.6]
mk-en	LER (%)	[28.8, 32.6]	[26.8, 30.5]	[25.2, 28.8]	[0.0, 0.0]
	BLEU (%)	[28.6, 31.0]	[30.7, 32.3]	[29.1, 31.5]	[30.9, 33.3]
eu-es	LER (%)	[43.6, 48.8]	[38.8, 44.2]	[40.9, 46.2]	[0.0, 0.0]
	BLEU (%)	[10.1, 12.0]	[10.6, 12.6]	[10.3, 12.2]	[11.5, 13.5]
en-es	LER (%)	[20.5, 24.9]	[15.1, 18.9]	**[10.4, 13.8]**	[0.0, 0.0]
	BLEU (%)	[21.5, 23.4]	[21.9, 23.8]	**[22.2, 24.1]**	[22.8, 24.7]

Table 4: LER and BLEU scores with 95% confidence intervals for the reference systems on the test corpora. The `max-ent` system has been trained using fractional counts. The results in bold face show statistically significant improvements for the maximum-entropy model compared to the TL model according to pair-bootstrap resampling.

Acknowledgements: We acknowledge support from the Spanish Ministry of Industry and Competitiveness through project Ayutra (TIC2012-32615) and from the European Commission through project Abu-Matran (FP7-PEOPLE-2012-IAPP, ref. 324414) and thank all three anonymous referees for useful comments on the paper.

References

Berger, A., Pietra, S. D., and Pietra, V. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Dagan, I. and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., ORegan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Ginestí-Rosell, M., Ramírez-Sánchez, G., Ortiz-Rojas, S., Tyers, F. M., and Forcada, M. L. (2009). Development of a free Basque to Spanish machine translation system. *Procesamiento de Lenguaje Natural*, (43):185–197.

Ide, N. and Véronis, J. (1998). Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–41.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proc. of the Conference on EMNLP*, pages 388–395.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proc. of the 10th MT Summit*, pages 79–86.

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proc. of the Annual Meeting of the ACL, demonstration session*.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mareček, D., Popel, M., and Žabokrtský, Z. (2010). Maximum entropy translation model in dependency-based MT framework. In *WMT '10 Proc. of the Joint 5th Workshop on SMT and MetricsMATR*, pages 201–206.

Melero, M., Oliver, A., Badia, T., and Suñol, T. (2007). Dealing with bilingual divergences in MT using target language n-gram models. In *Proc. of the METIS-II Workshop*, pages 19–26.

Nuhn, M. and Ney, H. (2014). Em decipherment for large vocabularies. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 759–764.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation. In *ACL-2002: 40th Annual meeting of the ACL*, pages 311–318.

Ravi, S. and Knight, K. (2011). Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 12–21. Association for Computational Linguistics.

Sánchez-Martínez, F., Pérez-Ortiz, J. A., and Forcada, M. L. (2008). Using target-language information to train part-of-speech taggers for machine translation. *Machine Translation*, 22(1-2):29–66.

Tyers, F. M. (2010). Rule-based Breton to French machine translation. In *Proc. of the 14th Annual Conference of the EAMT*, pages 174–181.

Tyers, F. M. and Alperen, M. S. (2010). SETimes: A parallel corpus of Balkan languages. In *Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages at the Language Resources and Evaluation Conference*, pages 1–5.

Tyers, F. M., Sánchez-Martínez, F., and Forcada, M. L. (2012). Flexible finite-state lexical selection for rule-based machine translation. In *Proc. of the 16th Annual Conference of the EAMT*, pages 213–220, Trento, Italy.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. (2005). Word-sense disambiguation for machine translation. In *Proc. of HLT Conference and Conference on EMNLP*, pages 771–778.

Zinovjeva, N. (2000). Learning sense disambiguation rules for machine translation. Master's thesis, Uppsala University.