Source-Language Dictionaries Help Non-Expert Users to Enlarge Target-Language Dictionaries for Machine Translation

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz

Transducens Research Group
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, Spain
{vmsanchez,mespla,japerez}@dlsi.ua.es

Abstract

In this paper, a previous work on the enlargement of monolingual dictionaries of rule-based machine translation systems by non-expert users is extended to tackle the complete task of adding both source-language and target-language words to the monolingual dictionaries and the bilingual dictionary. In the original method, users validate whether some suffix variations of the word to be inserted are correct in order to find the most appropriate inflection paradigm. This method is now improved by taking advantage from the strong correlation detected between paradigms in both languages to reduce the search space of the target-language paradigm once the source-language paradigm is known. Results show that, when the source-language word has already been inserted, the system is able to more accurately predict which is the right target-language paradigm, and the number of queries posed to users is significantly reduced. Experiments also show that, when the source language and the target language are not closely related, it is only the source-language part-of-speech category, but not the rest of information provided by the source-language paradigm, which helps to correctly classify the target-language word.

Keywords: Machine translation, Dictionaries, Non-expert users

1. Introduction

Rule-based machine translation (MT) systems heavily depend on explicit linguistic data such as monolingual dictionaries, bilingual dictionaries, grammars, etc. (Hutchins and Somers, 1992). Although some automatic acquisition is possible, collecting these data usually requires at some degree the intervention of linguists. In order to alleviate their work load, it could be interesting to open the door to a broader group of non-expert users who could enrich MT systems through the web.

We propose a novel method for enlarging some of these linguistic resources with the collaboration of non-expert users. In particular, the approach presented in this paper allows them to insert entries in the monolingual dictionaries and the bilingual dictionary of shallow-transfer rule-based MT systems. Notice that monolingual dictionaries encode the linguistic information of words in source language (SL) and target language (TL), while bilingual dictionaries contain mappings between SL and TL words.

In our system, non-expert users provide a SL word and its TL translation (for instance, *cars* and *coches*, for an English–Spanish MT system). Then, each word is inserted in the corresponding monolingual dictionary by assigning it to one of the paradigms defined for the corresponding language; these paradigms group regularities in inflection which are common to a set of words. The most appropriate paradigm is chosen by means of simple and easy yes/no questions which only require *speaker-level* understanding of the language. Basically, users are asked to validate whether the forms resulting from temptatively assigning some paradigms to the word to be inserted are correct forms of it. After that, the corresponding entry is inserted in the bilingual dictionary.

In a previous work we already tackled the enrichment of monolingual dictionaries by non-expert users (Esplà-Gomis et al., 2011); here, we extend it by exploiting the already inferred SL paradigm to reduce the search space of the TL paradigm (this is the main contribution of this paper) and adding an entry to the bilingual dictionary.

Note that non-expert users may get involved in the task in different ways:

- Users of an online translation service may be asked to provide additional information about some unknown words not being correctly translated because they are not present in the system dictionaries.
- A list of SL words may be uploaded to a crowdsourcing (Wang et al., 2010) platform by the developers of an MT system. Then, users willing to collaborate will provide the translation of each word and validate the different inflection alternatives proposed by our system. The evaluation of our approach has been carried out in this scenario, in the sense that a set of users has been provided with a list of words to be inserted in a dictionary. However, users have been directly recruited for this task (a crowdsourcing platform will be used in the future) and they have only inserted words in the TL monolingual dictionary.

In the experiments we have used the free/open-source rule-based MT system Apertium (Forcada et al., 2011), which is being currently used to build MT systems for a variety of language pairs.

The rest of the paper is organised as follows: section 2 outlines the most prominent related approaches and section 3 describes monolingual and bilingual dictionaries in the Apertium shallow-transfer rule-based MT system. Our approach is presented in section 4, then both human and automatic experiments carried out are described in section 5, and, finally, the results obtained are analysed and some con-

cluding remarks are presented in sections 6 and 7, respectively.

2. Related work

Similar approaches to ours can be divided in two groups: approaches which elicit the linguistic knowledge from users, and those which extract it from other resources.

Knowledge elicitation. Two of the more prominent works related to the elicitation of knowledge for building or improving MT systems are those by Font-Llitjós (2007) and McShane et al. (2002). The former proposes a strategy for improving both transfer rules and dictionaries by analysing the postediting process performed by a non-expert user through a special interface. McShane et al. (2002) design a complex framework to elicit linguistic knowledge from informants who are not trained linguists and use this information to build MT systems which translate into English; their system provides users with a lot of information about different linguistic phenomena to ease the elicitation task. Also, the probabilistic models of statistical MT systems can be estimated using active learning (Olsson, 2009) by asking users to translate the most informative sentences (Ambati et al., 2010).

Automatic extraction of resources. Many approaches have been proposed to deal with the automatic acquisition of linguistic resources for MT, mainly, transfer rules and dictionaries, even for the specific case of the Apertium platform (Caseli et al., 2006; Sánchez-Martínez and Forcada, 2009). The automatic identification of morphological rules (a problem for which paradigm identification is a potential resolution strategy) has also been subject of many recent studies (Monson, 2009; Creutz and Lagus, 2007; Goldsmith, 2010; Walther and Nicolas, 2011).

Novelty. Our work introduces some novel elements compared to previous approaches:

1. Unlike the Avenue formalism used in the work by Font-Llitjós (2007), the MT system we have used is a *pure* transfer-based one in the sense that a single translation is generated and no language model is used. Therefore, we are interested in the unique right answer and assume that an incorrect paradigm cannot be assigned to a new word.

2. Bartusková and Sedlácek (2002) also present a tool for semi-automatic assignment of words to declination patterns; their system is based on a decision tree with a question in every node. Their proposal, however, focuses on nouns and is aimed at experts because of the technical nature of the questions.

3. Our approach is addressed to non-experts, and, therefore, the answer to as few as possible simple questions is our main source of information . Font-Llitjós (2007) already anticipated the advisability of incorporating an active learning mechanism in her transfer rule refinement system, asking the user to validate different translations deduced from the initial hypothesis. However, this active learning approach has not yet been undertaken. Unlike the work by McShane et al. (2002), we want to relieve users of acquiring linguistic skills.

4. Our work focuses on identifying the paradigm which could be assigned to a word, a task more restrictive than decompounding a word into morphemes. The technique defined by Monson (2009) tolerates some errors in the final output.

3. The Apertium shallow-transfer rule-based MT system

This section describes in more detail the Apertium shallow-transfer rule-based machine translation system, and how data is encoded in its monolingual and bilingual dictionaries.

3.1. Overview

The process carried out by shallow-transfer systems can be split into three different steps:

1. Analysis of the SL text to build a SL intermediate representation, which is based on *lexical forms* consisting of lemma, part-of-speech category and morphological inflection information of the words in the input sentence. The SL monolingual dictionary is used to convert each SL *inflected word form* (IWF), i.e. a word as it is found in the text (for instance, *cars*), into a SL *lexical form* (*car, noun, plural*).

2. Transfer from that SL intermediate representation to a TL intermediate representation. The bilingual dictionary converts SL lexical forms (*car, noun, plural*) into TL lexical forms (in the case of Spanish, *coche, noun, masculine, plural*). Shallow-transfer rules, which are out of the scope of this paper, perform additional operations such as reordering and agreement on sequences which would not be correctly translated word-for-word using only the bilingual dictionary.

3. Generation of the final translation (*coches*) from the TL intermediate representation (*coche, noun, masculine, plural*) using the TL monolingual dictionary.

3.2. Monolingual dictionaries

Monolingual dictionaries have two types of data: *paradigms*, that group regularities in inflection, and *word entries*, represented by a stem and a paradigm. For instance, the paradigm assigned to many common English verbs, indicates that by adding *-ing* to the stem, the gerund is obtained; by adding *-ed*, the past is obtained; and so on. Paradigms make easier the management of dictionaries in two ways: by reducing the quantity of information that needs to be stored, and by simplifying revision and validation thanks to the explicit encoding of regularities in the dictionary. For example, describing the inflection of a verb by giving its stem (for instance, *wait*) and inflection model ("it is conjugated as") is safer than writing all the possible conjugated forms one by one. Once the most frequent paradigms in a dictionary are defined, entering a new word is generally limited to writing the stem and choosing an inflection paradigm. Our system helps to assign new words to existing paradigms in both monolingual dictionaries by efficiently interrogating the user.

3.3. Bilingual dictionaries

Bilingual dictionaries in Apertium contain the relationships between the lexical forms in SL and TL. The paradigm corresponding to each lexical form included in the bilingual dictionary can be easily obtained from the word entries in the monolingual dictionaries. In the resulting relationship between SL and TL paradigms, we observed that, usually, only a reduced set of TL paradigms correspond to a SL paradigm and only a smaller subset of them appear in a relatively high amount of entries related to the SL paradigm. This observation suggests that knowing the paradigm of a SL word may help to choose the best paradigm of its TL counterpart, which is the hypothesis we will test in the experiments.

In order to statistically confirm the observed close relationship between SL and TL paradigms, we may estimate the conditional probability $p(p_i^{TL}|p_j^{SL})$ as the number of entries whose SL paradigm is p_j^{SL} and whose TL paradigm is p_i^{TL} divided by the number of entries whose SL paradigm is p_j^{SL}. For a particular SL paradigm p_j^{SL}, the conditional entropy $H(p^{TL}|p_j^{SL})$ gives us an idea about the uncertainty of the TL paradigm of a word once we know that the paradigm of its SL equivalent is p_j^{SL}. If T is the set of TL paradigms, it is computed as:

$$H(p^{TL}|p_j^{SL}) = -\sum_{i \in T} p(p_i^{TL}|p_j^{SL}) \cdot \log p(p_i^{TL}|p_j^{SL})$$

We computed this entropy for all the SL paradigms of Apertium Catalan–Spanish and English–Spanish linguistic data, and the resulting histograms are shown in figures 1 and 2, respectively.

In both cases, the TL paradigms corresponding to the translation of most SL paradigms present a value of entropy under 0.5,[1] which confirms the strong correlation between the paradigms. Note that, the proportion of SL paradigms whose related TL paradigms have an entropy under 0.5 is higher in the Catalan–Spanish dictionaries, probably because they are more closely related than English–Spanish.

4. Method

As the methodology to insert entries in the monolingual dictionaries and the bilingual dictionary is built upon the previous work by Esplà-Gomis et al. (2011), a brief description of it follows before presenting the main contribution of this paper.

4.1. Baseline algorithm

Let $P = \{p_i\}$ be the set of paradigms in a monolingual dictionary. Each paradigm p_i defines a set of suffixes[2] $F_i = \{f_{ij}\}$ which are appended to stems to build new IWFs, along with some additional morphological information. Given a *stem/paradigm* pair c composed of a stem t and a paradigm p_i, the *expansion* $I(t, p_i)$ is the set of possible IWFs resulting from appending each of the suffixes in p_i to t. For instance, an English dictionary may contain a paradigm p_i with suffixes $F_i = \{\epsilon$,*-s*, *-ed*, *-ing*$\}$ (ϵ denotes the empty string), and the stem *want* assigned to p_i; the expansion $I(\text{want}, p_i)$ consists of the set of IWFs *want*, *wants*, *wanted* and *wanting*. We also define a *candidate stem* t as an element of $\Pr(w)$, the set of possible prefixes of a particular IWF w.

Given a new IWF w to be added to a monolingual dictionary, our objective is to find both the candidate stem $t \in \Pr(w)$ and the paradigm p_i which expand to the largest possible set of IWFs which are correct forms of w. To that end, our method performs the three tasks described below. It is worth noting that in this work we assume that all the paradigms for the words in the language are already included in the dictionary.

Paradigm detection. To detect the set of paradigms which may produce the IWF w and their corresponding stems we use a *generalised suffix tree* (GST) (McCreight, 1976) containing all the possible suffixes included in the paradigms in P. A list L is built containing all the candidate stem/paradigm pairs compatible with the IWF to be added (which we will also refer to as candidate paradigms (CPs)). We will denote each of these candidates with c_n.

The following example illustrates this stage of our method. Consider a simple dictionary with only three paradigms:

p_1, with F_1=$\{f_{11}$=ϵ, f_{12}=*-s*$\}$;
p_2, with F_2=$\{f_{21}$=*-y*, f_{22}=*-ies*$\}$;
p_3, with F_3=$\{f_{31}$=*-y*, f_{32}=*-ies*, f_{33}=*-ied*, f_{34}=*-ying*$\}$; and
p_4, with F_4=$\{f_{41}$=*-a*, f_{42}=*-um*$\}$.

Lets assume that a user wants to add the new IWF w=*policies* (corresponding to the noun *policy*) to the dictionary. The candidate stem/paradigm pairs which will be obtained after this stage are:

c_1=*policies*/p_1,
c_2=*policie*/p_1,
c_3=*polic*/p_2, and
c_4=*polic*/p_3.

Paradigm scoring. Once L is obtained, a *confidence score* is computed for each CP $c_n \in L$ using a large monolingual corpus C. The score considers the frequency of occurrence in the corpus of each IWF in each candidate c_n. In this way, candidates producing a set of IWFs which are more likely to appear in the corpus get higher scores.

Following our example, the IWFs for the different candidates would be:

$I(c_1)$=$\{$*policies*, *policiess*$\}$,
$I(c_2)$=$\{$*policie*, *policies*$\}$,
$I(c_3)$=$\{$*policy*, *policies*$\}$, and
$I(c_4)$=$\{$*policy*, *policies*, *policied*, *policying*$\}$.

Using a large monolingual English corpus C, IWFs *policies* and *policy* will be easily found, and the rest of them (*policie*, *policiess*, *policied* and *policying*) will not. Therefore, c_3 would obtain the highest score.

[1] An entropy of 0.5 corresponds to an uncertainty between that of a random variable with only one possible outcome (entropy 0) and that of a random variable with two equally likely outputs (entropy 1)

[2] Although it can be easily adapted to deal with prefix inflection, this methodology was originally designed to work with suffixes.

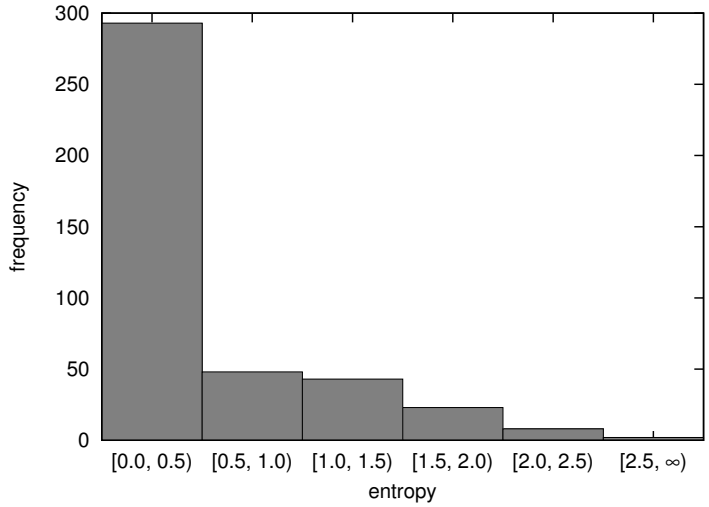

Figure 1: Histogram representing the value of the conditional entropy (see section 2) of the random variable representing the TL paradigm assigned to a word given the paradigm of its SL equivalent (according to the bilingual dictionary) for the Apertium Catalan–Spanish linguistic data. The total number of paradigms in the Catalan monolingual dictionary is 417. See section 5 for more details on the data used to obtain this chart.

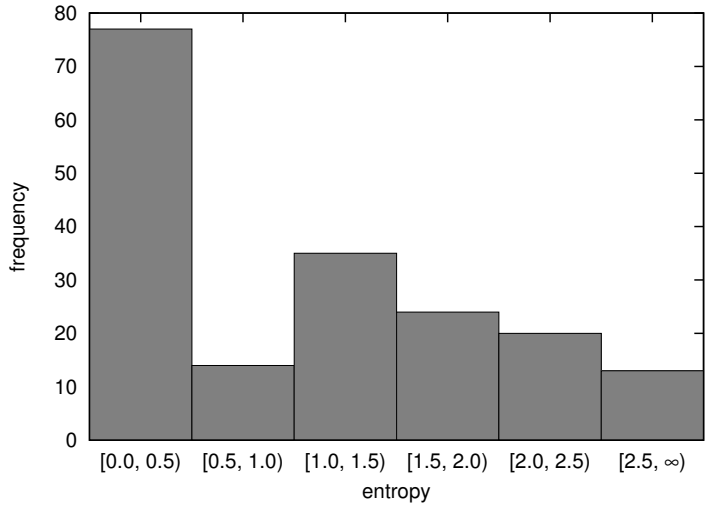

Figure 2: Histogram representing the value of the conditional entropy (see section 2) of the random variable representing the TL paradigm assigned to a word given the paradigm of its SL equivalent (according to the bilingual dictionary) for the Apertium English–Spanish linguistic data. The total number of paradigms in the English monolingual dictionary is 183. See section 5 for more details on the data used to obtain this chart.

User interaction. Finally, the best candidate is chosen from L by querying the user about a reduced set of the IWFs for some of the CPs $c_n \in L$. To do so, our system firstly sorts L in descending order using the confidence score previously computed. Then, users are asked (following the order in L) to confirm whether some of the IWFs in each expansion are correct forms of w. In this way, when an IWF w' is presented to the user

- if it is accepted, all $c_n \in L$ for which $w' \notin I(c_n)$ are removed from L;
- if it is rejected, all $c_n \in L$ for which $w' \in I(c_n)$ are removed from L.

In order to minimize the amount of yes/no questions asked, users are iteratively asked to validate the IWF from the first CP in L which is present in the minimum number of other CPs in L. If the confidence score is accurate, it is very likely to be accepted, and, consequently, the number of IWFs discarded would be maximized. This process is repeated until only one candidate remains in L. Note that this method cannot distinguish between candidates which produce the same set $I(c_n)$: in the experiments, they are considered as a single candidate.

4.2. Inserting words in the Apertium dictionaries

Given a SL IWF and its TL translation (for instance, *cars* and *coches* when the SL is English and the TL is Spanish), both provided by a non-expert user, our methodology for dictionary enrichment involves the following three steps:

Inserting the source-language word in the source-language monolingual dictionary. This step is performed by following the method by Esplà-Gomis et al. (2011), which has just been described.

Inserting the target-language word in the target-language monolingual dictionary by exploiting correlations between paradigms. The main contribution of this paper belongs to this step. The TL word is inserted in the TL monolingual dictionary following a method based on the one used in the SL monolingual dictionary (Esplà-Gomis et al., 2011) featuring an improved confidence score which takes advantage of the strong correlation which exists between paradigms in SL and TL (see section 3.3). In particular, the original confidence score of each TL CP p_i^{TL} is multiplied by the conditional probability $p(p_i^{TL}|p_j^{SL})$, where p_j^{SL} is the paradigm of the SL equivalent, in order to take into account the SL paradigm information. A simple smoothing is applied: when the value of one of the two factors is zero, it is replaced by the lowest non-zero value among all the candidate paradigms divided by 10.

Inserting an entry in the bilingual dictionary. As the SL and TL lemmas and the inflection information can be straightforwardly derived from the SL and TL paradigms, the corresponding bilingual dictionary entry can be added without any additional user interaction.

5. Experimental settings

Since the method for inserting an entry in the SL monolingual dictionary has already been evaluated (Esplà-Gomis et al., 2011), our experimental set-up is focused on studying the impact of the information provided by the SL paradigm when inferring the TL paradigm. We assume that the paradigm in SL is already known and focus on inserting the information of the translation in the TL monolingual dictionary.

We have used the Apertium Catalan–Spanish[3] and English–Spanish[4] language pairs and we have chosen a Spanish Wikipedia dump[5] as the monolingual corpus to compute the confidence scores.

Human evaluation has been carried out to test the impact of the information provided by the SL paradigm in closely related language pairs (Catalan–Spanish), while automatic evaluation has been designed to study how the performance changes when languages are not so closely related.

5.1. Human evaluation

The main goal of the human evaluation is to assess whether the information provided by the SL paradigm effectively reduces the amount of words users have to classify and whether, as a consequence of being queried about a lower amount of words, the amount of errors made by users is reduced and the TL paradigms are more accurately chosen. The basic idea under our evaluation strategy is to choose a set of common words from the Apertium Spanish monolingual dictionary, remove them from the dictionary, and ask a group of non-expert users to insert them using our method. Spanish acts as TL and Catalan as SL. The SL paradigms of the words from the test set can be easily obtained by checking the Catalan–Spanish bilingual dictionary and the Catalan monolingual dictionary. Consequently, a comparison can be performed between the addition of an entry using the SL paradigm information and not using it (Esplà-Gomis et al., 2011).

In order to build the test set, we firstly select the TL paradigms which meet the following restrictions:

- The lexical information they encode includes an open part-of-speech category. When creating the monolingual dictionary for a given language, words from closed part-of-speech categories constitute a small set which can be inserted by expert users.
- They have at least six words assigned in the dictionary.
- When removing the 5 most frequent words assigned to them, al least one inflection of one of the remaining words can be found in the monolingual corpus.

From the paradigms fulfilling the previous conditions, we choose the 30 paradigms whose words are most common in the monolingual corpus. From each of them, the 5 most common words are extracted and a set of 150 words is built. From each word, its most common inflection is chosen to obtain the final test word pool.

These 150 IWFs are divided into 4 subsets, one of them for each of the four non-expert human evaluators, introducing some redundancy which allows computing inter-annotator and intra-annotator agreement metrics. Each subset contains 50 IWFs and is built as follows:

- 30 IWFs extracted from the initial pool which are not shared with any other evaluator. Consequently, 120 IWFs are extracted from the pool in this way.
- 5 IWFs from that set are included twice, in order to compute intra-annotator agreement.
- Each pair of evaluators share 5 IWFs extracted from the remaining 30 IWFs in the word pool. Therefore, 15 IWFs from each evaluator subset are obtained in this way and, as there are 6 different evaluator pairs, the remaining 30 IWFs are used.

We asked each human evaluator to insert the words from his/her test set twice: firstly, using the enhanced method described previously, in which the information of the SL paradigm is taken into account, and secondly with the non-improved baseline system by Esplà-Gomis et al. (2011). A sentence from the monolingual corpus containing the IWF to be classified is shown to the user to ease the classification of homographs.

In addition to the inter-annotator and intra-annotator agreement metrics, the following evaluation scores were calculated:

[3]Revision 33900 in the Apertium SVN's trunk `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-es-ca`

[4]Revision 36247 in the Apertium SVN's trunk `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-en-es`

[5]`http://dumps.wikimedia.org/eswiki/20110114/eswiki-20110114-pages-articles.xml.bz2`

- success rate: percentage of words from the test set that have been tagged with the paradigm originally assigned to them in the monolingual dictionary;

- *average precision and recall*: precision (P) and recall (R) were computed as

$$P(c, c') = |I(c) \cap I(c')| \cdot |I(c)|^{-1},$$

$$R(c, c') = |I(c) \cap I(c')| \cdot |I(c')|^{-1},$$

  where c is the stem/paradigm pair chosen by our system and c' is the pair originally in the dictionary;

- average position of the right candidate in the sorted list of CPs; and

- average number of queries posed by our system to the non-expert users for every new word.

5.2. Automatic evaluation

As results of the human evaluation showed a strong correlation between the position of the right CP in the list L and the number of queries posed to users (see next section for more details), the first metric can be used to estimate the second one without human interaction. Therefore, we computed variation in the position of the right CP in L when using the SL paradigm information with a much bigger test set and with the English–Spanish and Catalan–Spanish language pairs, in order to detect if the performance of our method depends on how related are the languages involved.

We included in our test set the most common inflection in the monolingual corpus of all the word entries belonging to paradigms from open part-of-speech categories which have at least two word entries. For each IWF in the test set, its corresponding word entry was temporarily removed from the Apertium dictionary, the position of the correct CP in L was computed with the baseline algorithm (Esplà-Gomis et al., 2011) and with our new approach which takes into account the SL paradigm, and finally the removed word entry was restored to the dictionary so that it is available for the next word in the experiment.

One of the most important data a paradigm encodes is the part-of-speech category of the words belonging to it. We are interested in assessing whether an improvement in the accuracy of the confidence score when using the information of the SL paradigm is caused only by the correlation between the part-of-speech categories in SL and TL, or the rest of information provided by the paradigms is also useful. To do so in the context of automatic evaluation, we add a third strategy to calculate the position of the right paradigm in L: a modification of our approach in which the conditional probability $p(p_i^{TL}|p_j^{SL})$, calculated from the relative frequency of paradigms in the bilingual dictionary, is calculated assuming that, for a given language, all the paradigms generating the same part-of-speech category are grouped into a single one. In other words, the factor is the same one for all the candidate paradigms with the same part-of-speech category.

6. Results

6.1. Human evaluation

Regarding the quality of the annotators, table 1 shows the the pair-wise inter-annotator agreement computed using Cohen's kappa (Cohen, 1960), and table 2 the intra-annotator agreement for each annotator computed in the same way.

Annotator pair	κ (baseline)	κ (with SL paradigm)
A-B	0.76	0.76
A-C	0.74	0.74
A-D	0.71	0.71
B-C	0.76	0.76
B-D	1.00	1.00
C-D	1.00	1.00
average	0.83	0.83

Table 1: Inter-annotator agreement computed using Cohen's kappa for the 6 pairs of annotators.

Annotator	κ (baseline)	κ (with SL paradigm)
A	1.0	1.0
B	0.73	1.0
C	1.0	1.0
D	1.0	1.0
average	0.93	1.0

Table 2: Intra-annotator agreement computed using Cohen's kappa.

A kappa value between 0.6 and 0.8 is usually interpreted as *good agreement*, and when it ranges between 0.8 and 1.0 it is usually stated that there is a *very good agreement* between annotators. As all the values obtained fall in one of these ranges, we can conclude that each of the 4 annotators was quite consistent (high intra-annotator agreement) and that they agreed in their answers (high inter-annotator agreement), which ensures the trust of the results presented below.

Table 3 shows the value of the five evaluation metrics for our new approach and the baseline method. Confidence intervals were estimated with 95% statistical significance with a *t-test*.

Although neither success rate nor precision nor recall are improved when the information of the SL paradigm is included, a statistically significant improvement in the position of the correct paradigm in the list of CPs and in the number of queries posed to the evaluators is observed, which confirms that the information provided by the SL paradigm is valuable, at least in closely related languages.

In addition, it is worth noting the correlation detected between the position of the right paradigm in the initial sorted list of candidates and the average number of queries posed to the evaluators, as shown in figure 3.

The obtained success rate is high, and precision and recall are even higher, because those words which were assigned to incorrect paradigms, were assigned to paradigms producing similar IWFs. The superlative form of adjectives in Spanish was one of the main sources of errors: Apertium

System	success rate	P	R	initial position in L	# queries
baseline	$88\% \pm 5$	$94\% \pm 3$	$95\% \pm 3$	12.0 ± 2.0	6.1 ± 0.7
new method using SL paradigm	$87\% \pm 5$	$94\% \pm 2$	$98\% \pm 2$	2.3 ± 0.7	4.4 ± 0.4

Table 3: Success rate, precision, recall, position of the right paradigm in the initial sorted list of candidates and average number of queries posed to the evaluators (95% statistical significance) when inserting entries in the Apertium Spanish monolingual dictionary using the new method described in section 4 and a baseline in which the corresponding Catalan paradigm is not known.

contains paradigms for adjectives which have superlative form and for those which do not have it. Users often accepted the superlative form of an adjective which, according to the Apertium dictionary, does not have it.

6.2. Automatic evaluation

As shown in table 4, when extending the evaluation to the whole dictionary, a significant improvement in the position of the right candidate in L still happens, even in the case of a less related language pair such as Spanish–English.

However, results show that, for English–Spanish, the only information from the SL paradigm which helps to classify the TL word is the part-of-speech category. This can be explained by the fact that closely related languages share the inflection scheme (for instance, in Spanish and Catalan, nouns have gender and number) and most words keep their inflection features when they are translated (for instance, most masculine nouns whose plural is built by appending *-s* in Catalan, are also masculine and their plural is built in the same way in Spanish). Contrarily, inflection schemes are different in less related languages (such as English and Spanish) and, therefore, the inflection information encoded in paradigms is not useful.

7. Concluding remarks

We have extended our previous work on enlarging monolingual dictionaries of rule-based MT systems by non-expert users to tackle the complete task of adding both SL and TL forms of a word to the monolingual dictionaries and the bilingual dictionary. We have improved the original method by taking advantage from the strong correlation detected between paradigms in both languages. Results show that, when the SL word has already been inserted, a better ranking for the TL CPs can be obtained and, consequently, the amount of queries posed to users is significantly reduced as well. When SL and TL are not closely related, the only relevant information provided by the SL paradigm is the part-of-speech category.

In the near future, we plan to change how the different elements which make up the confidence score of candidate paradigms are aggregated. We will introduce a perceptron which allows tuning the weights of the different elements being combined. We will also study how to replace the heuristic algorithm which decides the IWFs users have to validate at each step with a a decision-tree-based one. In addition, we are currently developing a method to overcome one of the most important restrictions of our approach: it cannot distinguish between CPs which produce the same set of IWFs (but encode different lexical information). The new method consists basically in searching in a monolingual corpus the IWFs and deciding which is the most likely paradigm given the lexical information of the surrounding words.

8. Acknowledgements

This work has been partially funded by Spanish Ministerio de Ciencia e Innovación through project TIN2009-14009-C02-01 and by Generalitat Valenciana through grant ACIF/2010/174 from VALi+d programme.

9. References

V. Ambati, S. Vogel, and J. Carbonell. 2010. Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation, LREC 2010*.

D. Bartusková and R. Sedlácek. 2002. Tools for semi-automatic assignment of czech nouns to declination patterns. In *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, pages 159–164, London, UK. Springer-Verlag.

H. Caseli, M. Nunes, and M.L. Forcada. 2006. Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20:227–245.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

M. Creutz and K. Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process*, 4.

M. Esplà-Gomis, V. M. Sánchez-Cartagena, and J. A. Pérez-Ortiz. 2011. Enlarging monolingual dictionaries for machine translation with active learning and non-expert users. In *Proceedings of Recent Advances in Natural Language Processing*, pages 339–346, Hissar, Bulgaria, September.

A. Font-Llitjós. 2007. *Automatic improvement of machine translation systems*. Ph.D. thesis, Carnegie Mellon University.

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

J.A. Goldsmith, 2010. *The Handbook of Computational Linguistics and Natural Language Processing*, chapter Segmentation and morphology. Wiley-Blackwell.

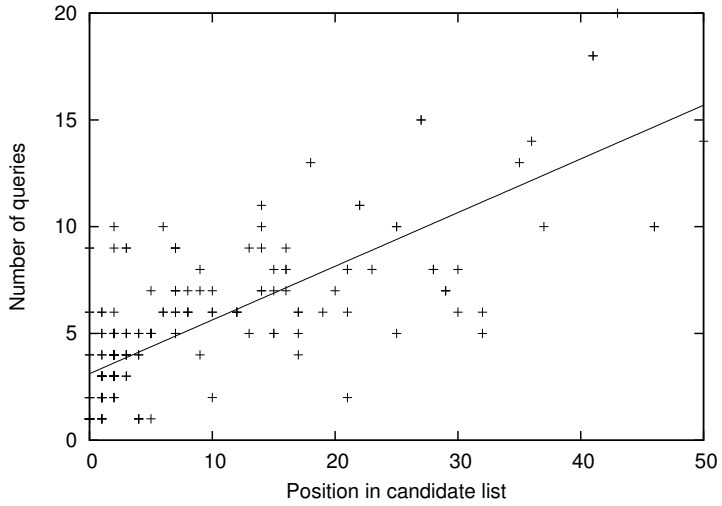

Figure 3: Correlation, obtained by least squares, between the position of the right candidate paradigm in the list of candidate paradigms (see section 4) and the number of queries for the baseline described in section 5.

Language pair	System	initial position in L
Catalan–Spanish	baseline	21.9 ± 0.2
	new method using SL part-of-speech category	15.1 ± 0.2
	new method using SL paradigm	13.2 ± 0.2
English–Spanish	baseline	26.1 ± 0.3
	new method using SL part-of-speech category	21.0 ± 0.3
	new method using SL paradigm	21.1 ± 0.3

Table 4: Average position of the right paradigm in the initial sorted list of candidates (95% statistical significance) when inserting, using the methods described in this paper, each entry in the Apertium Spanish monolingual dictionary from the Catalan–Spanish language pair and the Apertium Spanish monolingual dictionary from the English–Spanish language pair. Note that these experiments have been carried out without human interaction (see section 5.2).

W. J. Hutchins and H. L. Somers. 1992. *An introduction to machine translation*. Academic Press, London.

E.M. McCreight. 1976. A space-economical suffix tree construction algorithm. *Journal of the Association for Computing Machinery*, 23:262–272, April.

M. McShane, S. Nirenburg, J. Cowie, and R. Zacharski. 2002. Embedding knowledge elicitation and MT systems within a single architecture. *Machine Translation*, 17:271–305.

C. Monson. 2009. *ParaMor: From Paradigm Structure to Natural Language Morphology Induction*. Ph.D. thesis, Carnegie Mellon University.

F. Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. Technical report, School of Electronics and Computer Science, University of Southampton.

F. Sánchez-Martínez and M.L. Forcada. 2009. Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34:605–635.

G. Walther and L. Nicolas. 2011. Enriching morphological lexica through unsupervised derivational rule acquisition. In *Proceedings of the International Workshop on Lexical Resources*.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2010. Perspectives on crowdsourcing annotations for natural language processing. Technical report, School of Computing, National University of Singapore.