Enlarging Monolingual Dictionaries for Machine Translation with Active Learning and Non-Expert Users

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz
Transducens Research Group
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, Spain
{mespla,vmsanchez,japerez}@dlsi.ua.es

Abstract

This paper explores a new approach to help non-expert users with no background in linguistics to add new words to a monolingual dictionary in a rule-based machine translation system. Our method aims at choosing the correct paradigm which explains not only the particular surface form introduced by the user, but also the rest of inflected forms of the word. A large monolingual corpus is used to extract an initial set of potential paradigms, which are then interactively refined by the user through active machine learning. We show the results of experiments performed on a Spanish monolingual dictionary.

1 Introduction

Rule-based machine translation (MT) systems heavily depend on explicit linguistic data such as morphological dictionaries, bilingual dictionaries, grammars, and structural transfer rules (Hutchins and Somers, 1992). Although some automatic acquisition is possible, collecting these data usually requires in the end the intervention of domain experts (mainly, linguists) who master all the encoding and format details of the particular MT system. We should, however, open the door to a broader group of non-expert users who could collaboratively enrich MT systems through the web.

In this paper we present a novel method for enlarging the monolingual dictionaries in rule-based MT systems by non-expert users. An automatic process is first run to collect as much linguistic information as possible about the new word to be added to the dictionary and, after that, the resulting set of potential hypothesis is filtered by eliciting additional knowledge from non-experts with no linguistic background through *active learning* (Olsson, 2009; Settles, 2010), that is, by interactively querying the user in order to efficiently reduce the search space. As these users do not possess the technical skills which are usually required to fill in the dictionaries, this elicitation is performed via a series of simple and easy yes/no questions which only require *speaker-level* understanding of the language. Our method does not only incorporate to the dictionary the particular surface form introduced by the user (for example, *wants*), but it also discovers a suitable paradigm for the new word so that all the word forms of the corresponding lexeme and their morphological information (such as *wanted, verb, past* or *wanting, verb, gerund*) are also inserted.

This work focuses on monolingual dictionaries. These dictionaries have basically two types of data: *paradigms*, that group regularities in inflection, and *word entries*. The paradigm assigned to many common English verbs, for instance, indicates that by adding the ending *-ing*, the gerund is obtained. Paradigms make easier the management of dictionaries in two ways:

1. by reducing the quantity of information that needs to be stored, thereby creating more compact data structures, and

2. by simplifying revision and validation by describing the regularities in the dictionary; for example, describing the inflection of a verb by giving its stem 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 inflected word is generally limited to writing the stem and choosing an inflection paradigm. We show a semi-automatic method for the assignment of new words to the existing paradigms in a monolingual dictionary, which interrogates the user when it cannot automatically find enough evidence for unambiguously determining the correct paradigm. Note that as paradigms in MT usually contain morphological information (gender, noun, tense, etc.) on every inflected word form, our method also avoids

the user from identifying all these linguistic data.

In our experiments we will use the free/open-source rule-based MT system Apertium (Forcada et al., 2011). Apertium[1] is being currently used to build MT systems for a variety of language pairs. Every word is assigned to a paradigm in Apertium's monolingual dictionaries, and specific paradigms are defined for words with irregular forms. In addition, all the lexical information is included in the paradigms; as a result, there exist paradigms which only contain lexical information and do not add any suffix to the corresponding stem; the paradigm for the proper nouns is a good example of this.

Once a word and its corresponding translation have been added to the monolingual dictionaries of the source and target languages, respectively, of a MT system, the next step is to link both of them by adding the corresponding entry in the bilingual dictionary. How to adapt this task to non-experts is out of the scope of this paper and will be tackled in future works.

Social Translation. In spite of the vast amount of contents and collaboratively-created knowledge uploaded to the web during the last years, linguistic barriers still pose a significant obstacle to universal collaboration as they lead to the creation of "islands" of content, only meaningful to speakers of a particular language. Until *fully-automatic high-quality* MT becomes a reality, massive online collaboration in translation may well be the only force capable of tearing down these barriers (Garcia, 2009) and produce large-scale availability of multilingual information. Actually, this collaborative translation movement is happening nowadays, although still timidly, in applications such as Cucumis.org, OneHourTranslation.com or the Google Translator Toolkit[2].

The resulting scenario, which may be called *social translation*, will need efficient computer translation tools, such as reliable MT systems, friendly postediting interfaces, or shared translation memories. Remarkably, collaboration around MT should not only concern the postediting of raw machine translations, but also the creation and management of the linguistic resources needed by the MT systems; if properly done, this can lead to a significant improvement in the translation engines. Since as many hands as possible are necessary for the task, speakers that, in principle, do not have the level of technical know-how required to improve MT systems or manage linguistic resources must be involved, and, consequently, software that can make those tasks easier and elicit the knowledge of both experts and non-experts must be developed (Font-Llitjós, 2007; Sánchez-Cartagena and Pérez-Ortiz, 2010). This large-scale collaboration implies a change of paradigm in the way linguistic resources are managed and a series of conditions should hold in order to fully accomplish the goals of this social translation scenario (Pérez-Ortiz, 2010).

Knowledge Elicitation and Active Learning. 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 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 into English; their system provides users with a lot of information about different linguistic phenomena to ease the elicitation task. Ambati et al. (2010) show how to apply an active learning (Olsson, 2009) strategy to the configuration of a statistical machine translation.

Automatic Extraction of Resources. Many approaches have been proposed to deal with the automatic acquisition of linguistic resources for MT, mainly, transfer rules and bilingual 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), our MT system is a *pure* transfer-based one in the sense that a single translation is generated and no language model is used to score a set of possible candidate translations. Therefore, we are interested in the unique right answer and assume that an incorrect paradigm cannot be assigned to a new word.

[1] `http://www.apertium.org`

[2] `http://translate.google.com/toolkit`

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, including those who probably cannot define even vaguely what, for instance, an adverb is, but who can intuitively identify whether a particular word is correct under the rules for forming words in their language; therefore, the answer to as few as possible simple questions is our main source of information in addition to what an automated extraction method may deliver in a first step. 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 a set of morphemes. In the work by Monson (2009) some errors are tolerated in the final output of the system.

5. Our mid-term intention is to develop a system in line with the social translation principles which may be used to collaboratively build MT systems from scratch. This will also include the semi-automatic learning of the paradigms or the transfer rules which better serve the translation task, and which do not need necessarily correspond to the linguistically motivated ones.[3]

Outline of the Paper. The rest of the paper is organised as follows. Section 2 introduces our method for semi-automatic assignment of words to paradigms. A brief outline of the format used by the dictionaries of the Apertium MT system is given in section 3. Section 4 presents our experimental set-up and Section 5 discusses the results attained. The experiments performed pose some limitations in our approach or in the way in which data is currently represented in Apertium's dictionaries, which are discussed in section 6, together with some ideas on how to cope with them in future work. Finally, the paper ends with some conclusions.

[3]For example, a single inferred paradigm could group inflections for verbs like *wait* (ϵ, *-s*, *-ed*, *-ing*) and nouns like *waiter* (ϵ, *-s*), whereas an expert would probably write two different paradigms in this case.

2 Methodology

In this work we focus on languages which generate inflections by adding suffixes to the stems of words, as happens, for example, with Romance languages; our approach, however, could be easily adapted to inflectional languages based on different ways of adding morphemes. Let $P = \{p_i\}$ be the set of paradigms in a monolingual dictionary. Each paradigm p_i defines a set of suffixes $F_i = \{f_{ij}\}$ which are appended to stems to build new inflected word forms, along with some additional morphological information. The dictionary also includes a list of stems, each labelled with the index of a particular paradigm; the *stem* is the part of a word that is common to all its inflected variants. Given a *stem/paradigm pair* composed of a stem t and a paradigm p_i, the *expansion* $I(t, p_i)$ is the set of possible word forms resulting from appending all 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 word forms *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 word form w.

Given a new word form 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 morphologically correct inflections. To that end, our method performs three tasks: obtaining the set of all compatible stem/paradigm candidates which generate, among others, the word form w when expanded; giving a *confidence score* to each of the stem/paradigm candidates so that the next step is as short as possible; and, finally, asking the user about some of the inflections derived from each of the stem/paradigm candidates obtained in the first step. Next we describe the methods used for each of these three tasks.

It is worth noting that in this work we assume that all the paradigms for the words in the dictionary are already included in it. The situation in

which for a given word no suitable paradigm is available in the dictionary will be tackled in the future, possibly by following the ideas in related works (Monson, 2009).

2.1 Paradigm Detection

The first step for adding a word form w to the dictionary is to detect the set of *compatible* paradigms. To do so, we use a *generalised suffix tree* (GST) (McCreight, 1976) containing all the possible suffixes included in the paradigms in P. Each of these suffixes is labelled with the index of the corresponding paradigms. The GST data structure allows to retrieve the paradigms compatible with w by efficiently searching for all the possible suffixes of w; when a suffix is found, the prefix and the paradigm are considered as a candidate stem/paradigm pair. In this way, a list L of candidate stem/paradigm pairs is built; 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: f_{11}=ϵ, f_{12}=*-s*
p_2: f_{21}=*-y*, f_{22}=*-ies*
p_3: f_{31}=*-y*, f_{32}=*-ies*, f_{33}=*-ied*, f_{34}=*-ying*

Assume that a user wants to add the new word w=*policies* 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, c_4=*polic*/p_3

2.2 Paradigm Scoring

Once L is obtained, a *confidence score* is computed for each stem/paradigm candidate $c_n \in L$ using a large monolingual corpus C. One possible way to compute the score is

$$\mathrm{Score}(c_n) = \frac{\sum_{\forall w' \in I(c_n)} \mathrm{Appear_C}(w')}{\sqrt{|I(c_n)|}},$$

where $\mathrm{Appear_C}(w')$ is a function that returns 1 when the inflected form w' appears in the corpus C and 0 otherwise, and I is the expansion function as defined before. The square root term is used to avoid very low scores for large paradigms which include lot of suffixes.

One potential problem with the previous formula is that all the inflections in $I(c_n)$ are taken into account, including those that, although morphologically correct, are not very usual in the language and, consequently, in the corpus. To overcome this, $\mathrm{Score}(c_n)$ is redefined as

$$\mathrm{Score}(c_n) = \frac{\sum_{\forall w' \in I'_C(c_n)} \mathrm{Appear_C}(w')}{\sqrt{|I'_C(c_n)|}},$$

where $I'_C(c_n)$ is the difference set

$$I'_C(c_n) = I(c_n) \setminus \mathrm{Unusual}_C(c_n).$$

The function $\mathrm{Unusual}_C(c_n)$ uses the words in the dictionary already assigned to p_i as a reference to obtain which of the inflections generated by p_i are not usual in the corpus C. Let $T(p_i)$ be a function retrieving the set of stems in the dictionary assigned to the paradigm p_i. For each of the suffixes f_{ij} in F_i our system computes

$$\mathrm{Ratio}(f_{ij}, p_i) = \frac{\sum_{\forall t \in T(p_i)} \mathrm{Appear_C}(t f_{ij})}{|T(p_i)|},$$

and builds the set $\mathrm{Unusual}_C(c_n)$ by concatenating the stem t to all the suffixes f_{ij} with $\mathrm{Ratio}(f_{ij}, p_i)$ under a given threshold Θ.

Following our example, the following inflections for the different candidates will be obtained:

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

Using a large monolingual English corpus C, word forms *policies* and *policy* will be easily found; the other inflections (*policie*, *policiess*, *policied* and *policying*) will not be found. To simplify the example, assume that $\mathrm{Unusual}_C(c_n) = \emptyset$ for all the candidates; the resulting scores will be: $\mathrm{Score}(c_1)$=0.71, $\mathrm{Score}(c_2)$=0.71, $\mathrm{Score}(c_3)$=1.41, $\mathrm{Score}(c_4)$=1.

2.3 Active Learning Through User Interaction

Finally, the best candidate is chosen from L by querying the user about a reduced set of the inflections for some of the candidate paradigms $c_n \in L$. To do so, our system firstly sorts L in descending order by $\mathrm{Score}(c_n)$. Then, users are asked to confirm whether some of the inflections in each expansion are morphologically correct (more precisely, whether they *exist* in the language); the only possible answer for these questions is *yes* or *no*. In this way, when an inflected word form 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.

Note that c_1, the best stem/paradigm pair according to Score, may change after updating L. Questions are asked to the user until only one single candidate remains in L. In order to ask as few questions as possible, the word forms shown to the user are carefully selected. Let $G(w', L)$ be a function giving the number of $c_n \in L$ for which $w' \in I(c_n)$. We use the value of $G(w', L)$ in two different phases: *confirmation* and *discarding*.

Confirmation. In this stage our system tries to find a suitable candidate c_n, that is, one for which all the inflections in $I(c_n)$ are morphologically correct. In principle, we may consider that the inflections generated by the best candidate c_1 in the current L (the one with the highest score) are correct. Because of this, the user is asked about the inflection $w' \in I(c_n)$ with the lowest value for $G(w', L)$, so that, in case it is accepted, a significant part of the paradigms in L are removed from the list. This process is repeated until

- only one single candidate remains in L, which is used as the final output of the system; or
- all $w' \in I(c_1)$ are generated by all the candidates remaining in L, meaning that c_1 is a suitable candidate, although there still could be more suitable ones in L.

If the second situation holds, the system moves on to the *discarding* stage.

Discarding. In this stage, the system has accepted c_1 as a possible solution, but it needs to check whether any of the remaining candidates in L is more suitable. Therefore, the new strategy is to ask the user about those inflections $w' \notin I(c_1)$ with the highest possible value for $G(w', L)$. This process is repeated until

- only c_1 remains in L, and it will be used as the final output of the system; or
- an inflection $w' \notin I(c_1)$ is accepted, meaning that some of the other candidates is better than c_n.

If the second situation holds, the system removes c_1 from L and goes back to the *confirmation* stage.

For both *confirmation* and *discarding* stages, if there are many inflections with the same value for $G(w', L)$, the system chooses the one with higher $\mathrm{Ratio}(f_{ij}, p_i)$, that is, the most usual in C.

It is important to remark that this method cannot distinguish between candidates which generate the same set $I(c_n)$. In the experiments, they have considered as a single candidate.

In our example, the ordered list of candidates will be $L = (c_3, c_4, c_1, c_2)$. Choosing the inflection in $I(c_3)$ with the smaller value for $G(w', L)$ the inflection *policy*, which is only generated by two candidates, wins. Hopefully, the user will accept it and this will make that c_1 and c_2 be removed from L. At this point, $I(c_3) \subset I(c_4)$, c_3 is suitable and, consequently, the system will try to discard c_4. Querying the user about any of the inflections in $I(c_4)$ which is not present in $I(c_3)$ (*policied* and *policying*) and getting user rejection will make the system to remove c_4 from L, confirming c_3 as the most suitable candidate.

3 Monolingual Dictionaries in Apertium

A small example follows to show how a simple entry is encoded in the English Apertium's monolingual dictionary. A paradigm named *par123* to be used in English nouns with singular ending in *-um* which change it to *-a* to form the plural form will be defined in XML as follows:

```
<pardef n="par1">
  <e><p>
    <l>um</l>
    <r>um<s n="n"/><s n="sg"/></r>
  </p></e>
  <e><p>
    <l>a</l>
    <r>um<s n="n"/><s n="pl"/></r>
  </p></e>
</pardef>
```

Now, the words *bacterium/bacteria* and *datum/data* will be defined as follows:

```
<e lm="bacterium">
  <i>bacteri</i>
  <par n="par123"/>
</e>
<e lm="datum">
  <i>dat</i>
  <par n="par123"/>
</e>
```

The part inside the *i* element contains the stem of the lexeme, which is common to all inflected forms, and the element *par* refers to the assigned paradigm. In this case, *bacterium* will be analysed into `bacterium<n><sg>` and *bacteria* into `bacterium<n><pl>`.

It is also possible to create entries in the dictionaries consisting of two or more words if these words are considered to build a single *translation unit*. Dictionaries may also contain *nested*

paradigms used in other paradigms (for instance, paradigms for enclitic pronoun combinations are included in all Spanish verb paradigms).

It is clear that it may be hard for non-experts to incorporate new entries to the dictionaries unless methods, like the one proposed in this paper, exist to conveniently elicit their language knowledge.

4 Experiments

The aim of the experiments is to asses, in a realistic scenario, whether our semi-automatic methodology is valid to find out, for a given word, its most suitable paradigm. Therefore, a group of people has been told to add a set of words to a monolingual dictionary using our methodology. For this task, we chose the Apertium Spanish monolingual dictionary from the language pair Spanish–Catalan. First, the dictionary was filtered to remove

- word entries belonging to a closed part-of-speech category: when building a monolingual dictionary from scratch, words from closed categories are usually included first, since they are very frequent in source texts;
- word entries assigned to a paradigm which only contains an empty suffix: these paradigms usually define proper nouns, which may be identified using other methods;
- multi-word units, which are out of the scope of this paper;
- prefix inflection entries: as our methodology is designed to deal with suffix inflection, the only entry found in the dictionary with prefix inflection was discarded;
- redundant paradigms, which generate the same inflections with the same lexical information and are, therefore, equivalent.

A test set was created with words extracted from the filtered dictionary. Firstly, a stem assigned to each of the paradigms p_i with $1 < |T(p_i)| < 10$ was added. To build a more realistic test set, we chose one more stem from those paradigms p_i with $10 \leq |T(p_i)|$ in order to have more words assigned to very common paradigms. Then, we obtained, for each pair stem/paradigm, all the possible word forms and included the most common ones into the test set using the $\mathrm{Ratio}(f_{ij}, p_i)$ value. In this way, we obtained 226 words: 106 extracted from the first group of paradigms and 120 from the second one. Obviously, the stems from which we obtained the words included in the test set were removed from the dictionary.

Then, the test set was split into 10 subsets, and each subset was assigned to a different human evaluator. Each evaluator in an heterogeneous group of non-experts was then asked to introduce each of the words in their test set using our system. Experiments were run using the filtered dictionary and a word list obtained from the Spanish Wikipedia dump[4] as the monolingual corpus C.

The different evaluation metrics obtained from the human evaluation process are:

- *success rate*: number of words from the test set that have been tagged with the paradigm assigned to them in the original Apertium dictionary. This is the most straightforward metric to evaluate our methodology;
- *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. Confidence intervals were estimated with 99% statistical confidence with a *t-test*;
- *average number of questions*: average number of questions made by our system for each word in the test set;
- *average number of initial paradigms*: the average number of compatible paradigms initially found as possible solutions in the first stage of our method.

The value of the threshold Θ used to compute the set $\mathrm{Unusual}_C(c_n)$ defined in Section 2 was 0.1.

Finally, an alternative approach without user interaction was designed as a baseline so that the impact of active learning could be better evaluated. The baseline consists of directly choosing the first element in the list L as the most suitable candidate. The average position of the right candidate in L has also been computed.

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

5 Results and Discussion

We evaluated our approach and computed the results following the metrics depicted in Section 4. The average number of initial candidates detected by our approach was 56.4; this metric was specially high for verbs, whereas it was much lower for nouns and adjectives. The average number of questions asked to the users by the active learning approach for the test set was 5.2, which is reasonably small considering that the 56.4 initial paradigms on average and that the average position of the right candidate in L was 9.1. Figure 1 shows an histogram representing the position of the right candidate in the initial list L for each word in the test set. We also observed that, in average, users needed around 30 seconds in average to find the paradigm of each word in the test set.

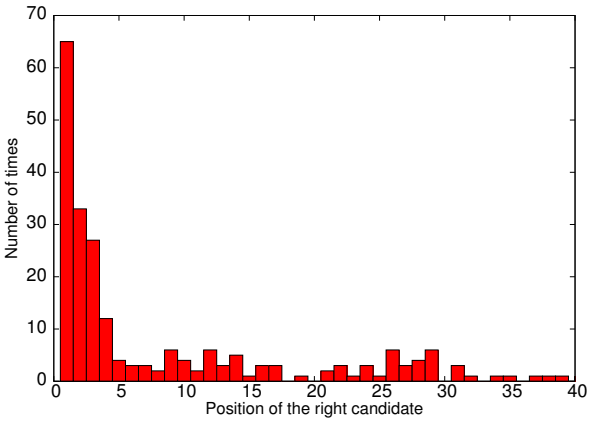

Figure 1: Histogram representing the distribution of the position of the right candidate in the initial list of candidates L for each word in the test set.

We obtained a success rate of 72.9% for the active learning approach with a precision of $P = 87\% \pm 5$ and a recall of $R = 87\% \pm 5$. These results stress the fact that those words which were assigned to incorrect paradigms, were assigned to paradigms generating similar inflections. These results are clearly better than those obtained by the baseline approach, with a success rate of 28.9%, a precision of $P = 70.3\% \pm 6$ and a recall of $R = 62.77\% \pm 7$.

Taking a closer look at the results, we observed some relevant causes for the errors. On the one hand, we detected human errors for words which should have been accepted but were rejected or vice-versa. These mistakes, caused by a lack of knowledge of the users (for example, about accentuation rules), should be taken into account in the future; they could be solved, for instance, by using *reinforcement questions* or combining the answers of different users for the same or similar words. Moreover it could be possible to give a kind of *confidence score* to the paradigms in the dictionary based on how frequently words are incorrectly assigned to them.

We also observed that most of the words which were not assigned to the expected paradigm were verbs. Spanish morphological rules allow multiple concatenations of enclitic pronouns at the end of verbs. In many occasions, users rejected forms of verbs with too many enclitic pronouns or for which some concrete enclitics had no semantic sense. This happens because, in order to reduce the number of possible paradigms, Apertium's dictionaries can assign some words to existing paradigms which are a superset of the correct one; since the included semantically incorrect word forms will never occur in a text to translate, this, in principle, may be safely done.

6 Limitations and Work Ahead

In this paper we have described a system for interactively enlarging dictionaries and selecting the most suitable paradigm for new words. Our preliminary experiments have brought to light several limitations of our method which will be tackled in the future.

Detection of lexical information. One of the most important limitations of our approach is that, as already commented in Section 2, candidate paradigms generating the same $I(c_n)$ set cannot be distinguished. This situation usually holds when the expansions of two different stem/paradigm pairs are equal but the lexical information in each paradigm is different. For example, in Spanish two different paradigms may contain the same suffixes F=$\{\epsilon$, -$s\}$ although one of them generates substantives and the other one generates adjectives.

We have started to explore a method to semi-automatically obtain this lexical information. A statistical part-of-speech tagger may be used to obtain initial hypothesis about the lexical properties of a word w; this information could then be refined by querying users with complete sentences in which w plays different lexical roles.

Lack of suitable paradigms. Our approach assumes that all the paradigms for a particular language are already included in the dictionary, but it could be interesting to have a method to also add new paradigms. The work by Monson (2009) could be a good start for the new method.

Other improvements. We plan to improve our approach by using simple statistical letter models of bigrams or trigrams to discard candidates generating morphologically unlikely word forms, or by using additional information in the scoring stage, such as word context, number of occurrences, etc.

7 Conclusions

We have shown an active learning method for adding new entries to monolingual dictionaries. Our system allows non-expert users with no linguistic background to contribute to the improvement of RBMT systems. The Java source code for the tool described in this paper is published[5] under an open-source license.

Acknowledgements

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

References

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

Dita Bartusková and Radek 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.

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

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

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

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martnez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*. doi: 10.1007/s10590-011-9090-0.

Ignacio Garcia. 2009. Beyond translation memory: Computers and the professional. *The Journal of Specialised Translation*, 12:199–214.

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

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

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

Marjorie McShane, Sergei Nirenburg, James Cowie, and Ron Zacharski. 2002. Embedding knowledge elicitation and MT systems within a single architecture. *Machine Translation*, 17:271–305.

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

Fredrik 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.

Juan Antonio Pérez-Ortiz. 2010. Social Translation: How Massive Online Collaboration Could Take Machine Translation to the Next Level. In *Second European Language Resources and Technologies Forum: Language Resources of the Future, FlarenetNet Forum 2010*, pages 64–65, Barcelona, Spain.

Víctor M. Sánchez-Cartagena and Juan Antonio Pérez-Ortiz. 2010. Tradubi: open-source social translation for the Apertium machine translation platform. In *Open Source Tools for Machine Translation, MT Marathon 2010*, pages 47–56.

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

Burr Settles. 2010. Active learning literature survey. Technical report, Computer Sciences Technical Report 1648, University of WisconsinMadison.

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

[5] `https://apertium.svn.sourceforge.net/svnroot/apertium/branches/apertium-dixtools-paradigmlearning`