English-Spanish Large Statistical Dictionary of Inflectional Forms

Grigori Sidorov[1,2], **Alberto Barrón-Cedeño**[2], **Paolo Rosso**[2]

[1]Center for Computing Research,
Instituto Politécnico Nacional
(National Polytechnic (Technical) Institute),
Mexico
[2]Natural Language Engineering Lab. - ELiRF
Department of Information Systems and Computation,
Universidad Politécnica de Valencia,
Spain
sidorov@cic.ipn.mx, {lbarron, prosso}@dsic.upv.es
(http://www.dsic.upv.es/grupos/nle)

Abstract

The paper presents an approach for constructing a weighted bilingual dictionary of inflectional forms using as input data a traditional bilingual dictionary, and not parallel corpora. An algorithm is developed that generates all possible morphological (inflectional) forms and weights them using information on distribution of corresponding grammar sets (grammar information) in large corpora for each language. The algorithm also takes into account the compatibility of grammar sets in a language pair; for example, verb in past tense in language L normally is expected to be translated by verb in past tense in Language L. We consider that the developed method is universal, i.e. can be applied to any pair of languages. The obtained dictionary is freely available. It can be used in several NLP tasks, for example, statistical machine translation.

1. Introduction

In a bilingual dictionary, a word w in a language L is linked to all its potential translations w' in a language L'. In a traditional bilingual dictionary a head word is usually a lemma, i.e. a morphologically normalized word form. Its translation very often is also a lemma or a set of possible lemmas. This is a typical situation, see below the discussion of more complex situations when the translation is a word combination.

Statistical bilingual dictionaries are a special type of bilingual dictionaries that contain, for a pair of words $\{w, w'\}$ how likely is that w' be a valid translation of w, i.e., $p(w', w)$. These dictionaries usually contain word forms (not lemmas) on both sides (Och and Ney, 2003) and are widely exploited in various Natural Language Processing (NLP) applications, such as Statistical Machine Translation (Brown et al., 1990) and Cross-Language Information Retrieval (Levow et al., 2005) as well as Cross-Language Plagiarism Detection (Barrón-Cedeño et al., 2008).

Most of the statistical bilingual dictionaries are estimated by considering parallel corpora on the basis of alignment methods such as the well known IBM M1 (Brown et al., 1990). The translation probabilities $p(w', w)$ are learned empirically from the parallel textual data. However, according to Zipf law, the appearance of every lemma and word form in such parallel texts is not guaranteed. Therefore, the generation of a dictionary that contains the entire collection of word paradigms (i.e., all possible word forms for each lemma) for large vocabularies is practically impossible using parallel texts.

This fact becomes particularly relevant (in a negative way) if the dictionary is exploited in order to process texts on a topic different from those covered in the training corpus. The lack of general vocabulary and, of course, all potential inflectional forms may cause the breakdown of the entire process. Therefore, it is necessary to generate dictionaries, or at least dictionary seeds, with a rich content in terms of vocabulary and inflectional forms.

In this paper, we describe the following achieved goals: (i) generation of a bilingual dictionary that includes a complete variation of words inflections, i.e. all possible word forms for each lemma for languages L and L' (though any pair of languages can be considered, in this case we considered L = English and L' = Spanish); (ii) estimation of the translation probabilities of each pair of word forms on the basis of monolingual frequencies of grammar classes in large corpora.

The rest of the paper is structured as follows. In Section 2 we discuss how the head words and translation equivalents can be represented with lemmas or word combinations. Afterwards, in Section 3. we describe how the dictionary is generated and how inflectional correspondences are weighted. Finally, Section 4 draws some conclusions and discusses potential applications of the dictionaries generated with the proposed method..

2. Translation Equivalents Represented with Word Combinations

As mentioned before, the typical situation in a bilingual dictionary is the presence of a head word (lemma) in L and one or several translation equivalents (lemmas) in L'. Sometimes, the situation is more complex when the translation equivalents are represented by a combination of words. A question arises for our task: how a word that is not a head word should be treated in the word combinations? I.e. should they be considered also as possible translation equivalents?

In some specialized dictionaries, like terminological dictionaries, even a head word can be represented as a word combination, for example, *concept album - disco monográfico*. The simplest solution that we adapt in this case is the usage of some heuristics or partial syntactic analysis for determining the syntactic structure of the word combination and then processing only the top head word.
Translations of the head word often are lemmas as well. Nevertheless, in this case it is much more frequent having translation equivalents represented as word combinations. The same considerations as above are applied. For the moment, we use just the top head word (*nucleus*) of the word combination.
Generally speaking, translation equivalents can be either a generalization, or, more often, a specification of the translated word. This specification can be whether (i) a set of adjectives that depend on the head word; (ii) a word combination where the translation equivalent is a lemma and the depending words have morphological forms that correspond to its government pattern; or (iii) a subordinate clause. It is desirable to treat somehow the dependant words because they represent part of the meaning of the word in the other language. However, they cannot be treated in the same way as the head word because these words are not translation equivalents of the head word in the other language but only specifiers.
All these considerations represent an interesting problem for further investigation.

3. Generation of the Dictionary

For the achievement of the beforementioned goals, we developed a corresponding algorithm for the pair of languages {English, Spanish}. The algorithm is divided into two main steps:

1. (i) morphological generation: creation of a complete list of word forms for a list of translation equivalents in each language; and
2. (ii) calculation of translation probabilities: estimation of the probabilities $p(w' \mid w)$ for all $w' \in L'$, $w \in L$.

As a word form can correspond to various lemmas it has several sets of possible inflectional correspondences in the other language.

3.1. Morphological Generation

Morphological generation is based on a list of bilingual correspondences. Its source was a traditional bilingual dictionary containing about 30,000 entry words and including around 64,000 translations. In order to generate the English and Spanish word forms we used the morphological dictionaries available in the FreeLing package (Atserias et al., 2006). The idea is to consider not only those pairs included in a traditional translation dictionary, but also all the possible inflectional forms of each pair of words "source word – translation word(s)". The generation process is summarized in Fig. 1.
An example of the list of inflectional forms obtained for a word form in English is presented in Table 1. It includes a word form of the verb *to take*, in this case *took*, with its valid translations into Spanish word forms.

Table 1: Example of generation for the word form "*took*" (grammar information is given for illustration purposes only).

Possible Spanish Translation	$p(w' \mid took)$
tomó_VMIS3S0	0.3016546
tomaba_VMII3S0;VMII1S0	0.2752902
tomaban_VMII3P0	0.0800329
tomaron_VMIS3P0	0.0670665
tomé_VMIS1S0	0.0528457
tomamos_VMIS1P0;VMIP1P0	0.0494479
tomase_VMSI3S0;VMSI1S0	0.0424848
tomara_VMSI3S0;VMSI1S0	0.0424848
tomasen_VMSI3P0	0.0121436
tomaran_VMSI3P0	0.0121436
tomar_VMN0000	0.0113312
toma_VMM02S0;VMIP3S0	0.0091485
tomábamos_VMII1P0	0.0087611
tomado_VMP00SM	0.0059050
tomaste_VMIS2S0	0.0044491
toman_VMIP3P0	0.0033597
tomabas_VMII2S0	0.0033013
tomando_VMG0000	0.0023740
tomada_VMP00SF	0.0019706
tomásemos_VMSI1P0	0.0017167
tomáramos_VMSI1P0	0.0017167
tomo_VMIP1S0	0.0014987
tomados_VMP00PM	0.0014060
tome_VMSP3S0;VMSP1S0;VMM03S0	0.0011019
tomadas_VMP00PF	0.0008767
tomases_VMSI2S0	0.0007872
tomaras_VMSI2S0	0.0007872
tomaría_VMIC3S0;VMIC1S0	0.0006075
tomará_VMIF3S0	0.0005070
tomen_VMSP3P0;VMM03P0	0.0004208
tomas_VMIP2S0	0.0004094
tomabais_VMII2P0	0.0002844
tomasteis_VMIS2P0	0.0002235
tomarán_VMIF3P0	0.0001992
tomaseis_VMSI2P0	0.0001874
tomarais_VMSI2P0	0.0001879
tomarían_VMIC3P0	0.0001489
tomemos_VMSP1P0;VMM01P0	0.0001304
tomes_VMSP2S0	0.0001065
tomaré_VMIF1S0	0.0000988
tomaremos_VMIF1P0	0.0000946
tomarás_VMIF2S0	0.0000477
tomaríamos_VMIC1P0	0.0000433
tomarens_VMSF3P0	0.0000413
tomáremos_VMSF1P0	0.0000410
tomareis_VMSF2P0	0.0000410
tomáis_VMIP2P0	0.0000320
tomad_VMM02P0	0.0000258
tomarías_VMIC2S0	0.0000136
toméis_VMSP2P0	0.0000111
tomaréis_VMIF2P0	0.0000062
tomare_VMSF3S0;VMSF1S0	0.0000017
tomares_VMSF2S0	0.0000015
tomaríais_VMIC2P0	0.0000008

Algorithm 1. Input: $Dict_{en-es}$

Initialize the set $T_{en,es}$
For each pair $\{en, es\} \in Dict_{en-es}$
$en_l = lemma(en)$;
$es_l = lemma(es)$
$F[en_l] \leftarrow word\_forms(en_l, English)$
$F[es_l] \leftarrow word\_forms(es_l, Spanish)$
Add $F[en_l] \times Fes_l$ to $T_{en,es}$
Return: $T_{en,es}$

Figure 1: Morphological generation algorithm. $T_{en,es}$ = set of generated translation pairs; $Dict_{en-es}$ = input bilingual dictionary; $lemma(x)$ function that generates the lemma of the word x; $word\_forms(x)$ function that generates all word forms for the lemma x.

3.2. Calculation of Translation Probabilities

A problem arises how to assign the probability for each translation $p(w', w)$. We use the idea that the probability of a word form is proportional to the distribution of the corresponding grammar sets in a large corpus. We use the term *grammar set* as part of a complete grammar paradigm for a given lemma. We consider that a paradigm is a well-structured table where all word forms can be placed, and grammar set characterizes each cell of this table. In this case, for example, *take* as a noun has two possible grammar sets (*Singular* and *Plural*), and *take* as a verb has at least four grammar sets that correspond to *take, takes, took, taken*. The exact number of grammar sets depends on how many cells we postulate for a verb in its paradigm for English language. An important point here is that we count probabilities for *take* as a noun and *take* as a verb separately and independently, because they have different grammar paradigms.

We considered frequencies of grammar sets for English and Spanish. The frequency distribution of English grammar sets (cf. Table 2) was estimated by considering a version of the WSJ corpus.[1] The frequency distribution of Spanish grammar sets (cf. Table 3) was calculated using a corpus marked with grammar information.[2] The English and Spanish corpora contain about 950,000 and 5.5 million word forms, respectively; a sufficient amount of words for our purposes. The frequencies included in Tables 2 and 3 give us the possibility to assign probabilities to word forms according to the proportion of their grammar sets (grammar information) in the corpora.

Though in theory a word form w can be translated by any word form w' with some probability, in most of the cases, these translations are highly improbable. In other words, *a priori* not every w can be likely translated into any w'.

We use a similarity measure between grammar classes in languages L and L'. For example, a noun in singular is more likely to be translated into a noun in singular than in plural. It is not expected that a verb in present tense would be translated into a verb in past tense. In order to calculate this similarity measure we developed an algorithm for our specific language pair, though the majority of its steps and conditions are rather universal. Indeed, the algorithm is applied to the language pair where Spanish has relatively rich morphology, while English has relatively poor morphological system. So, we consider that the algorithm is rather universal and can be applied to any pair of languages. If one of the languages has a reduced morphology, like, for example, Chinese, the algorithm still will be working for the other language. If we have a pair of two Chinese-like languages, then the algorithm will produce trivial results in cases of one- to-one translations. Another interesting question here is: how the algorithm will work for agglutinative languages, for example, Turkish, where a word has hundreds of grammar forms. Our first

Table 2: Distribution of English grammar classes.

Frequency	Grammar	Frequency	Grammar
163935	NN	11997	MD
121903	IN	10801	POS
114053	NNP	10241	PRP$
101190	DT	4042	JJR
75266	JJ	3275	RP
73964	NNS	3087	NNPS
38197	RB	2887	WP
37493	VBD	2625	WRB
32565	VB	2396	JJS
29462	CC	2175	RBR
26436	VBZ	555	RBS
24865	VBN	441	PDT
21357	PRP	219	WP$
18239	VBG	117	UH
15377	VBP		

Table 3: Distribution of Spanish grammar classes.

Frequency	Grammar	Frequency	Grammar
779175	SPS00	81613	DA0MP0
350406	NCFS000	78262	AQ0MS0
343046	NCMS000	...	
219842	DA0MS0	3	VSSI2P0
201115	CC	3	VSSF3P0
197969	RG	3	VASF1S0
187499	DA0FS0	3	VAM02P0
170729	NP00000	3	AQXMS0
147818	NCMP000	2	VASI2P0
137967	CS	2	VAIS2P0
136731	VMN0000	2	P02CP000
116310	NCFP000	2	AQXFS0
106492	VMIP3S0	2	AQXCP0
93495	PR0CN000	1	VSSF2S0
88735	AQ0CS0	1	VSM02S0
81613	DA0MP0	1	VSM02P0
78262	AQ0MS0	1	VMSF3S0
73092	DI0MS0	1	VASF3P0
71255	VMP00SM	1	VAM01P0
67882	P0000000	1	VAIC2P0
64774	AQ0FS0	1	PX2MP0P0
59394	VMIS3S0	1	PX1FP0S0
57661	DI0FS0	1	PT0FS000
56185	RN	1	AQXMP0
52512	VMII1S0	1	AQACP0

[1] Data obtained by José-Miguel Benedí, Universidad Politécnica de Valencia; jbenedi@dsic.upv.es

[2] http://www.lsi.upc.edu/∼nlp/web/

impression is that the algorithm will provide proper results if a large enough corpus is available. The problem is that some elements that are expressed lexically in synthetic languages will be expressed grammatically in agglutinative languages. This fact turns us back to the problem of translation using word combinations on both sides (see Section 2.).

The algorithm returns a boolean value[3] indicating if the grammar class in language L is compatible with the grammar class in language L'. The algorithm includes verification of conditions like those mentioned above, e.g., `if (English word is <Noun, Sg> and Spanish word is <Noun, Sg>) then return true`, etc.

Still, we would like to comment on one language-specific decision that we made: given an English verb, we consider that English past participle and gerund are compatible with practically any Spanish verb form in indicative. This decision is made because such verb forms are often part of compound tenses (perfect tenses and continuous tenses). For the same reason, Spanish participle and gerund are considered compatible with any English verb form.

In those cases where the grammar classes are incompatible, a very low probability is assigned to the translation into the implied word form. We use a threshold ϵ for the sum of all "incompatible" forms. Thus, all "compatible" word forms are equally distributed with the value of $1 - \epsilon$ (this will be weighted by the grammar distribution later). For instance, consider that, for a set of potential translations $p(w', w)$, the set of word forms w' consist of two compatible and three incompatible forms. The probability associated to the compatible forms will be $p(w', w) = (1 - \epsilon)/2$, and for the incompatible forms, it will be $p(w', w) = \epsilon/3$.[4]

Once we obtain the similarity estimations for all possible translations of word forms from one language into another on the basis of compatibility of the corresponding grammar classes, we follow on with the estimation of probabilities based on grammar distribution. This distribution establishes how likely is the appearance of the word form w with the given grammar class GC. It is calculated as:

$$g_d(w_{GC}) = \frac{freq(GC)}{\sum_{GC \in L} freq(GC)} \quad (1)$$

This estimation is based on the relative frequency of the grammar class GC in a significatively large corpus of language L. This process is carried on separately for each language. Finally, the translation probability for a pair (w, w') is estimated as follows:

$$p(w', w) = g_d w' \cdot g_d w \cdot \varrho(w' \mid w) \quad (2)$$

Note that we are interested in the probability of translations of a word form. If several grammar tags correspond to only one word form (for instance, consider the form *toma* in Table 1), the probability of the corresponding translation is the result of the sum of probabilities associated to each grammar tag, i.e.:

$$\varrho(w' \mid w) = \sum_{GC} p(w'_{GC} \mid w) \quad (3)$$

Finally, in order to obtain actual probabilities, the obtained values are scaled such that:

$$\sum_{w'} p(w' \mid w) = 1 \quad (4)$$

The generated dictionary is applicable to both translation directions.

4. Final Remarks

The produced statistical bilingual dictionary, currently available for English-Spanish translation, represents a useful resource for various NLP applications.[5] It was generated on the basis of a traditional bilingual dictionary (without using parallel texts) and includes translations of all possible combinations of inflectional forms between the implied languages. Translation probabilities are assigned according to the distributions of grammar forms in large corpora of the corresponding languages.

For the moment, we started from the English side and generated the dictionary on the basis of the valid Spanish translations. It seems that the dictionary generated from the other side will be equivalent (making corrections to the changes of the list of possible translations). We leave for future work its exact estimation.

As current work we are exploiting this resource for the generation of statistical dictionaries on the basis of alignment methods such as the IBM M1. For instance, Giza++ includes the option to provide a dictionary to the input of the alignment-based estimation of a statistical dictionary. We expect that the amount of noisy word equivalents in the resulting dictionary decreases.

5. Acknowledgements

The research work of the first author has been partially supported by the National Polytechnic Institute (SIP, COFAA, SIP grant 20090772), Mexican government (CONACYT/SNI), and the program *Estancias en la UPV de investigadores de prestigio* PAID-02-09 num. 3143. The research work of the second author has been partially supported by the CONACyT-Mexico 192021 grant. We thank the TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 research project. We also thank anonymous reviewers for their important comments.

6. References

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell Gonzáles, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library. In *Proceedings of the Fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA*, Genoa, Italy. http://www.lsi.upc.edu/ nlp/freeling.

[3] In future work, we plan to use real instead of boolean values.

[4] The value of ϵ must be estimated empirically. In this case we considered $\epsilon = 0.025$.

[5] The dictionary is freely available at http://users.dsic.upv.es/grupos/nle/downloads.html

Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, and Alfons Juan. 2008. On Cross-lingual Plagiarism Analysis Using a Statistical Model. In Benno Stein, Efstathios Stamatatos, and Moshe Koppel, editors, *Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 9–13, Patras, Greece.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vicent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-Based Techniques for Cross-Language Information Retrieval. *Information Processing and Management: Special Issue on Cross-Language Information Retrieval*, 41(3):523–547.

Frank Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51. See also http://www.fjoch.com/GIZA++.html.