English-Spanish Large Statistical Dictionary of Inflectional Forms

Grigori Sidorov[1], Alberto Barrón[2], and Paolo Rosso[2]

[1]Center for Computing Research (CIC),
National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz s/n, Zacatenco,
DF, 07738, Mexico
sidorov@cic.ipn.mx

[2]Polytechnic University of Valencia,
Valencia, Spain
{prosso, lbarron}@dsic.upv.es

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. morphologically normalized word form. Another term for lemma is grammar *type*, as opposed to grammar *token*, that is also known as *word form*. Translations of a head word also are lemmas. An exception from this situation is translation by word combination that has syntactic government, when the head word is lemma and the dependant word has the morphological form that corresponds to the government pattern.

There exists a special type of bilingual dictionaries called statistical bilingual dictionaries. These dictionaries usually contain word forms (not lemmas) on both sides (Och and Ney, 2003). This kind of dictionaries is widely used in various NLP applications like statistical machine translation, cross-language information retrieval (in particular, multilingual plagiarism detection), cross-language text clustering, among other tasks.

Most of the statistical bilingual dictionaries are obtained by considering parallel corpora on the basis of methods such as IBM-1 model, i.e. the probabilities of word forms are learned empirically from the parallel textual data. So, it is practically impossible to consider the entire vocabulary of a language including all potential inflectional forms (i.e., all possible word forms for each lemma), because it is not guaranteed that all lemmas and all word forms will occur in training texts. Especially, a problem arises if we are interested in the automatic processing of a text on different topic. The lack of general vocabulary and, of course, all potential inflectional forms can 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.

On the other hand, not all word forms have equal probabilities to appear in texts. It can be estimated by the calculation of the individual frequencies of word forms, but for this we need an enormous corpus, and still we cannot be sure in our results due to the Zipf law. Still, we leave to future work to evaluate the possibility to add individual frequencies to our corpus. Another possibility is to estimate these frequencies on the basis of frequencies of corresponding grammar classes.

In this paper, we achieve the following goals:

1. Generate a bilingual dictionary that includes a complete variation of words inflections, i.e. all possible word forms for each lemma (i.e., all tokens for each type), for both languages.

2. Estimate the translation probabilities of each pair of word forms on the basis of monolingual frequencies of grammar classes in large corpora.

3. Make a preliminary analysis how a list of anchored translations (i.e., a statistical dictionary generated *ad hoc*) affects the estimation of a statistical dictionary generated by some statistical machine translation system, e.g., Giza++.

The result of our processing is a large bilingual dictionary of inflectional forms with assigned probabilities that is a resource that can be used in various NLP applications.

Generation of the Dictionary

For achievement of the above mentioned goals we developed a corresponding algorithm for English and Spanish language pair. Algorithm contains two main steps: generation of a complete list of word forms for a given lemma in each language and estimation of the probabilities of all possible translations of the given word form. Note that a word form can correspond to various lemmas, and, thus, have several sets of possible inflectional correspondences in the other language.

Morphological generation

Still, first of all, we should base our generation on some list of bilingual correspondences. We used a traditional bilingual dictionary available in Internet as this source. It contains about 30,000 entry words and gives about 64,000 translations. For the moment we start from the English side and generate the dictionary on the basis of the Spanish translations. It seems that the dictionary generated from the other side – from Spanish to English – will be equivalent (making correction to the changes of the list of possible translations). We leave for future work the exact estimation.

For generation of the English and Spanish word forms we used the morphological dictionaries available in FreeLing package (FreeLing, 2009).

For each English lemma, the algorithm took all its word forms, and for each English word form it took all possible Spanish translations (according to the bilingual dictionary), lemmatized them, and for each Spanish lemma generated all possible word forms. An example resulting list for one English word form is presented in Table 1. Other word forms of the verb "*take*" with their corresponding Spanish word forms are also represented in the dictionary.

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

	English	Spanish	Probability
1.	took_VBD;	tomó_VMIS3S0;	0,3016546
2.	took_VBD;	tomaba_VMII3S0;VMII1S0;	0,2752902
3.	took_VBD;	tomaban_VMII3P0;	0,0800329
4.	took_VBD;	tomaron_VMIS3P0;	0,0670665
5.	took_VBD;	tomé_VMIS1S0;	0,0528457
6.	took_VBD;	tomamos_VMIS1P0;VMIP1P0;	0,0494479
7.	took_VBD;	tomase_VMSI3S0;VMSI1S0;	0,0424848
8.	took_VBD;	tomara_VMSI3S0;VMSI1S0;	0,0424848

	English	Spanish	Probability
9.	took_VBD;	tomasen_VMSI3P0;	0,0121436
10.	took_VBD;	tomaran_VMSI3P0;	0,0121436
11.	took_VBD;	tomar_VMN0000;	0,0113312
12.	took_VBD;	toma_VMM02S0;VMIP3S0;	0,0091485
13.	took_VBD;	tomábamos_VMII1P0;	0,0087611
14.	took_VBD;	tomado_VMP00SM;	0,0059050
15.	took_VBD;	tomaste_VMIS2S0;	0,0044491
16.	took_VBD;	toman_VMIP3P0;	0,0033597
17.	took_VBD;	tomabas_VMII2S0;	0,0033013
18.	took_VBD;	tomando_VMG0000;	0,0023740
19.	took_VBD;	tomada_VMP00SF;	0,0019706
20.	took_VBD;	tomásemos_VMSI1P0;	0,0017167
21.	took_VBD;	tomáramos_VMSI1P0;	0,0017167
22.	took_VBD;	tomo_VMIP1S0;	0,0014987
23.	took_VBD;	tomados_VMP00PM;	0,0014060
24.	took_VBD;	tome_VMSP3S0;VMSP1S0;VMM03S0;	0,0011019
25.	took_VBD;	tomadas_VMP00PF;	0,0008767
26.	took_VBD;	tomases_VMSI2S0;	0,0007872
27.	took_VBD;	tomaras_VMSI2S0;	0,0007872
28.	took_VBD;	tomaría_VMIC3S0;VMIC1S0;	0,0006075
29.	took_VBD;	tomará_VMIF3S0;	0,0005070
30.	took_VBD;	tomen_VMSP3P0;VMM03P0;	0,0004208
31.	took_VBD;	tomas_VMIP2S0;	0,0004094
32.	took_VBD;	tomabais_VMII2P0;	0,0002844
33.	took_VBD;	tomasteis_VMIS2P0;	0,0002235
34.	took_VBD;	tomarán_VMIF3P0;	0,0001992
35.	took_VBD;	tomaseis_VMSI2P0;	0,0001879
36.	took_VBD;	tomarais_VMSI2P0;	0,0001879
37.	took_VBD;	tomarían_VMIC3P0;	0,0001484
38.	took_VBD;	tomemos_VMSP1P0;VMM01P0;	0,0001305
39.	took_VBD;	tomes_VMSP2S0;	0,0001068
40.	took_VBD;	tomaré_VMIF1S0;	0,0000986
41.	took_VBD;	tomaremos_VMIF1P0;	0,0000947
42.	took_VBD;	tomarás_VMIF2S0;	0,0000473
43.	took_VBD;	tomaríamos_VMIC1P0;	0,0000433
44.	took_VBD;	tomaren_VMSF3P0;	0,0000410
45.	took_VBD;	tomáremos_VMSF1P0;	0,0000410
46.	took_VBD;	tomareis_VMSF2P0;	0,0000410
47.	took_VBD;	tomáis_VMIP2P0;	0,0000328
48.	took_VBD;	tomad_VMM02P0;	0,0000256
49.	took_VBD;	tomarías_VMIC2S0;	0,0000131
50.	took_VBD;	toméis_VMSP2P0;	0,0000112
51.	took_VBD;	tomaréis_VMIF2P0;	0,0000067
52.	took_VBD;	tomare_VMSF3S0;VMSF1S0;	0,0000017
53.	took_VBD;	tomares_VMSF2S0;	0,0000015
54.	took_VBD;	tomaríais_VMIC2P0;	0,0000008

Grammar distribution analysis

Now let us discuss the problem of assignment of probabilities to each pair of word forms. For estimation of probabilities we used the idea that the probability of a word form is proportional to the distribution of the corresponding grammar class in some large corpus.

We took English POS data from (English POS frequency list, 2009). This data was obtained from a version of WSJ corpus. Spanish data was taken from a corpus marked with grammar information (Spanish frequency lists, 2009). English corpus contains about 950 thousand word forms, while Spanish corpus contains about 5,5 million word forms. For our purposes it is important that they are big enough. English data and the fragment of Spanish data are presented in Table 2 and Table 3. This data gives us the possibility to assign frequencies to word forms according to proportion of their grammar information in the corpora.

Table 2. Distribution of the Spanish grammar classes.

Frequency	Grammar
779175	SPS00
350406	NCFS000
343046	NCMS000
219842	DA0MS0
201115	CC
197969	RG
187499	DA0FS0
170729	NP00000
147818	NCMP000
137967	CS
136731	VMN0000
116310	NCFP000
106492	VMIP3S0
93495	PR0CN000
88735	AQ0CS0
81613	DA0MP0
78262	AQ0MS0
73092	DI0MS0
71255	VMP00SM
67882	P0000000
64774	AQ0FS0
59394	VMIS3S0
57661	DI0FS0
56185	RN
52512	VMII1S0
52272	DA0FP0
40541	VMIP3P0
...	
3	VSSI2P0
3	VSSF3P0
3	VASF1S0
3	VAM02P0
3	AQXMS0
2	VASI2P0
2	VAIS2P0
2	P02CP000
2	AQXFS0
2	AQXCP0
1	VSSF2S0
1	VSM02S0
1	VSM02P0
1	VMSF3S0
1	VASF3P0
1	VAM01P0
1	VAIC2P0
1	PX2MP0P0
1	PX1FP0S0
1	PT0FS000
1	AQXMP0
1	AQACP0

Table 3. Distribution of the English grammar classes.

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

Algorithm for calculation of probabilities

Now let us describe the assignment of probabilities to the pairs of word forms. First of all, let us note that *a priori* not every word form can be likely translated by any of the other word form, for example, noun in singular is much more likely to be translated with another noun in singular than in plural. The verb in present tense is not expected to be translated by the verb in past, etc. Still, these are expected values that sometimes can not take place due to decisions of a translator. For taking into account this fact, we developed a measure of similarity between grammar forms in English and Spanish. We added there some obvious considerations mentioned above. For the moment, we assigned to English past participle and gerund probabilities to match practically any Spanish verb forms in indicative because they are part of compound tenses (perfect tenses and continuous tenses), as well as we assign high probabilities of the similar Spanish forms as they have the same function.

In case that our similarity measure returns zero, we assign very low probability to the word form. We used threshold of 0.025 for the sum of all “incompatible” forms, thus, all “compatible” word forms are distributed equally (this will be weighted by distribution later) with the value of 0.975. For example, if there are 2 compatible forms and 3 incompatible forms, then compatible forms will be assigned the value of 0.975/2 and incompatible forms of 0.025/3. The choice of this value is empirical. If several grammar tags correspond to a word form, then we sum the probabilities of each tag because finally we are interested in probability of a pair of words (word forms).

After the calculation of similarity of all possible word forms in the other language with the word form in the given language, we pass to the grammar distribution processing. We multiply each value of the form to its proportion in the corpus. It is done separately for each language.

At the next step, we multiply probabilities of each word form in the pair. E.g., if both word forms (English and Spanish) have the probability of 0.5, then the probability of a

pair is 0.5 * 0.5 = 0.25. Since each word form in each language is part of a set of all possible word forms, and, thus, has its probability according to grammar distribution, we prefer to take it into account. This makes our dictionary symmetrical and applicable to both languages. Still, strictly speaking, for a probability of the given translation, it is not necessary. We leave for future work to analyze what kind of dictionary is better.

Finally, we scale the values to match exactly the interval [0, 1].

Conclusions

We present a large bilingual dictionary of inflectional forms with assigned probabilities that is a resource that can be used in various NLP applications. The dictionary is generated starting from the bilingual dictionary (and not parallel texts) and contains all possible combinations of inflectional forms. The probabilities are assigned according to distributions of grammar forms in big corpora of the corresponding languages. We worked with English and Spanish language pair.

Our preliminary manual evaluation using GIZA++ shows that this dictionary reduces significantly probabilities of some improbable translations.

References

1. Franz Josef Och, Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29 (1), March 2003, pp. 19-51. See also http://www.fjoch.com/GIZA++.html
2. FreeLing. TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain, http://www.lsi.upc.edu/~nlp/freeling/. Consulted October 1, 2009.
3. Spanish frequency lists. TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain, http://www.lsi.upc.es/~nlp. Consulted October 1, 2009.
4. English POS frequency list. In: José Miguel Benedı. Métodos Estadísticos en Tecnologías del Lenguaje (2009-2010). www.dsic.upv.es/~jbenedi/docencia/metl/t3.pdf. Consulted October 1, 2009.