Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages

Ezeiza N., Alegria I., Arriola J.M., Urizar R.
Informatika Fakultatea
649 P.K Donostia E-20080
jibecran@si.ehu.es
http://ixa.si.ehu.es

Aduriz I.
UZEI
Aldapeta, 20.
Donostia E-20009
uzei@sarenet.es

Laburpena

Artikulu honetan metodo estokastiko eta erregeletan oinarritutako metodoen arteko konbinaketa euskarari aplikatzearen emaitzak aurkeztuko ditugu.Desanbiguazioan erabilitako metodoak Murrizpen Gramatika (CG) eta MULTEXT proiektuak garatutako HMMn oinarritutako etiketatzailea dira.
Euskara hizkuntza eranskaria izaki, hitz bakoitzari dagozkion irakurketa guztiak esleitzeko analizatzaile morfologikoa beharrezkoa da. Ondoren, CG erregelak informazio morfologiko guztiari aplikatzen zaizkio eta prozesu honek testuen anbiguotasuna gutxitzen du. Azkenik, geratutako etiketen artean bakarra hautatzeko MULTEXT proiektuko tresnak erabiltzen dira.
Metodo estokastikoa soilik erabiltzean, errore-tasa %14 ingurukoa da, baina etiketatzailearen doitasuna hitz ezezagunekin lexikoa aberastuz gero %2 hobe daitekeen arren. Metodo biak konbinatzen direnean, berriz, prozesu osoaren errore-tasa %3.5ekoa da. Ikasketarako corpusa nahikoa txikia dela, HMM eredua lehenengo mailakoa eta euskararako Murrizpen Gramatika oraindik ere garapen prozesuan dagoela kontuan izanik, gure ustez metodo konbinatu hau erabilita emaitza onak lor daitezke eta beste hizkuntza eranskarietarako bereziki egokia izan daiteke.

Resum

En aquest article presentem els resultats de la combinació de mètodes estocàstics i basats en regles aplicats a la desambiguació morfosintàctica de l'euskara. Els mètodes utilitzats per a la desambiguació són: les Gramàtiques de Restriccions (CG) i l'etiquetador basat en HMM del projecte MULTEXT.
El caràcter aglutinant de l'euskara fa necessari la utilització d'un analitzador morfològic per assignar a cada paraula totes les seves interpretacions. Les regles de CG s'apliquen utilitzant la informació morfològica completa i aquest procés redueix parcialment l'ambigüitat dels textos. A continuació, s'apliquen les eines de MULTEXT per escollir una única etiqueta.
Utilitzant només el mètode estocàstic la taxa d'error és aproximadament del 14%, encara que la precisió de l'etiquetador es pot incrementar en un 2% utilitzant les paraules desconegudes per enriquir el lèxic. En canvi, la combinació d'ambdós mètodes permet reduir l'error fins al 3.5%.
Tenint en compte que el corpus d'aprenentatge és bastant petit, que el model HMM és de primer ordre i que la Gramàtica de Restriccions de l'euskara està encara en fase de desenvolupament, creiem que els resultats del mètode combinat són bons i que la combinació de mètodes és especialment adequada per a llengües aglutinants.

Resumen

En este artículo presentamos los resultados de la combinación de métodos estocásticos y basados en reglas aplicados al euskara. Los métodos utilizados para la desambiguación son las Gramáticas de Restricciones (CG) y el etiquetador basado en HMM del proyecto MULTEXT.
Siendo el euskara una lengua aglutinante, será necesario un analizador morfológico para asignar a cada palabra todas sus interpretaciones. A continuación se aplican las reglas de CG utilizando toda la información morfológica y este proceso disminuye la ambigüedad de los textos. Por último, las herramientas de MULTEXT escogerán una única etiqueta.
Utilizando únicamente el método estocástico la tasa de error es de alrededor del 14%, aunque la precisión del etiquetador puede incrementarse en un 2% utilizando las palabras desconocidas para enriquecer el léxico. En cambio, combinando ambos métodos la tasa de error del proceso completo es del 3.5%. Teniendo en cuenta que el corpus de aprendizaje es bastante pequeño, que el modelo HMM es de primer orden y que la Gramática de Restricción del euskara está aún en fase de desarrollo, creemos el método combinado obtiene buenos resultados y puede ser adecuado para otras lenguas aglutinantes.

Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages

Ezeiza N., Alegria I., Arriola J.M., Urizar R.
Informatika Fakultatea
649 P.K Donostia E-20080
jibecran@si.ehu.es
http://ixa.si.ehu.es

Aduriz I.
UZEI
Aldapeta, 20.
Donostia E-20009
uzei@sarenet.es

Abstract

In this paper we present the results of the combination of stochastic and rule-based disambiguation methods applied to Basque language[1]. The methods we have used in disambiguation are Constraint Grammar formalism and an HMM based tagger developed within the MULTEXT project.
As Basque is an agglutinative language, a morphological analyser is needed to attach all possible readings to each word. Then, CG rules are applied using all the morphological features and this process decreases morphological ambiguity of texts. Finally, we use the MULTEXT project tools to select just one from the possible remaining tags.
Using only the stochastic method the error rate is about 14%, but the accuracy may be increased by about 2% enriching the lexicon with the unknown words. When both methods are combined, the error rate of the whole process is 3.5%. Considering that the training corpus is quite small, that the HMM model is a first order one and that Constraint Grammar of Basque language is still in progress, we think that this combined method can achieve good results, and it would be appropriate for other agglutinative languages.

Introduction

Based on the results of the combination of stochastic and rule-based disambiguation methods applied to Basque language, we will show that the results of the combination are significantly better than the ones obtained applying the methods separately.
As Basque is an agglutinative and highly inflected language, a morphological analyser is needed to attach all possible interpretations to each word. This process, which may not be necessary in other languages such as English, makes the tagging task more complex. We use MORFEUS, a robust morphological analyser for Basque developed at the University of the Basque Country (Alegria *et al.*, 1996). We present it briefly in section 1, in the overview of the whole system, the lemmatiser/tagger for Basque EUSLEM.
We have added to MORFEUS a lemma disambiguation process, described in section 2, which discards some of the analyses of the word based on statistical measures.
Another important issue concerning a tagger is the tagset itself. We discuss the design of the tagset in section 3.
In section 4, we present the results of the application of rule-based and stochastic disambiguation methods to Basque.
These results are deeply improved by combining both methods as explained in section 5.
Finally, we discuss some possible improvements of the system and future research.

1 Overview of the system

The disambiguation system is integrated in EUSLEM, a lemmatiser/tagger for Basque (Aduriz *et al.*, 1996). EUSLEM has three main modules:

- MORFEUS, the morphological analyser based on the two-level formalism. It is a robust and wide coverage analyser for Basque.
- the module that treats multiword lexical units. It has not been used in the experiments in order to simplify the process.
- the disambiguation module, which will be described in sections 5 and 6.

MORFEUS plays an important role in the lemmatiser/tagger, because it assigns every token all the morphological features. The most important functions are:

- incremental analysis, which is divided in

[1] This research has been supported by the Education Department of the Government of the Basque Country and the Interministerial Commision for Science and Technology.

three phases, using the two level formalism in all of them: 1) the standard analyser processes words according to the standard lexicon and standard rules of the language; 2) the analyser of linguistic variants analyses dialectal variants and competence errors[2]; and 3) the analyser of unknown words or guesser processes the remaining words.
- lemma disambiguation, presented below.

2 Lemma disambiguation

The lemma disambiguation has been added to the previously developed analyser for two main reasons:
- the average number of interpretations in unknown words is significantly higher than in standard words.
- there could be more than one lemma per tag. Since the disambiguation module won't deal with this kind of ambiguity, it has to be solved to lemmatise the text.

We use different methods for the disambiguation of linguistic variants and unknown words. In the case of linguistic variants we try to select the lemma that is "nearest" to the standard one according to the number of non-standard morphemes and rules. We choose the interpretation that has less non-standard uses.

	before	after
variants	2.58	2.52
unknown	13.1	6.21

Table 1- Number of readings.

In the case of unknown words, the procedure uses the following criteria:
- for each category and subcategory pair, leave at least one interpretation.
- assign a weight to each lemma according to the final trigram and the category and subcategory pair.
- select the lemma according to its length and weight –best combination of high weight and short lemma.

These procedures have been tested with a small corpus and the produced error-rate is 0.2%. This is insignificant considering that the average number of interpretations of unknown words decreases by 7, as shown in table 1.

3 Designing the tagset

The choice of a tagset is a critical aspect when designing a tagger. Before defining the tagset we have had to take some aspects into account: there was not any exhaustive tagset for automatic use, and the output of the morphological analyser is too rich and does not offer a directly applicable tagset.

While designing the general tagset, we tried to meet the following requirements:
- it had to take into account all the problems concerning ellipsis, derivation and composition (Aduriz *et al.*, 1995).
- in addition, it had to be general, far from *ad hoc* tagsets.
- it had to be coherent with the information provided by the morphological analyser.

Bearing all these considerations in mind, the tagset has been structured in four levels:
- in the first level, general categories are included (noun, verb, etc.). There are 20 tags.
- in the second level each category tag is further refined by subcategory tags. There are 48 tags.
- the third level includes other interesting information, as declension case, verb tense, etc. There are 318 tags in the training corpus, but using a larger corpus we found 185 new tags.
- the output of the morphological analysis constitutes the last level of tagging. There are 2,943 different interpretations in this training corpus, but we have found more than 9,000 in a larger corpus.

	ambiguity rate	tags/token
first	35.11%	1.48
second	40.68%	1.57
third	62.24%	2.20
fourth	64.42%	3.48

Table 2- Ambiguity of each level.

The morphological ambiguity will differ depending on the level of tagging used in each case, as shown in table 2.

4 Morphological Disambiguation

There are two kinds of methods for morphological disambiguation: on one hand, statistical methods need little effort and obtain very good results (Church, 1988; Cutting *et al.*, 1992), at least when applied to English, but when we try to apply them to Basque we encounter additional problems; on the other hand, some rule-based systems (Brill, 1992; Voutilainen *et al.*, 1992) are at least as good as statistical systems and are better adapted to free-order languages and agglutinative languages. So, we

[2] This module is very useful since Basque is still in normalisation process.

have selected one of each group: Constraint Grammar formalism (Karlsson *et al.*, 1995) and the HMM based TATOO tagger (Armstrong *et al.*, 1995), which has been designed to be applied it to the output of a morphological analyser and the tagset can be switched easily without changing the input text.

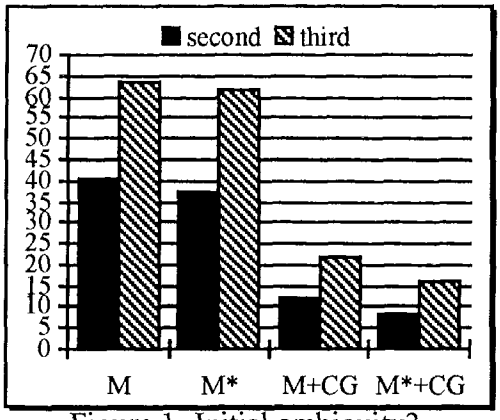

Figure 1- Initial ambiguity3.

We have used the second and third levels tagsets for the experiments and a small corpus –28,300 words– divided in a training corpus of 27,000 words and a text of 1,300 words for testing.

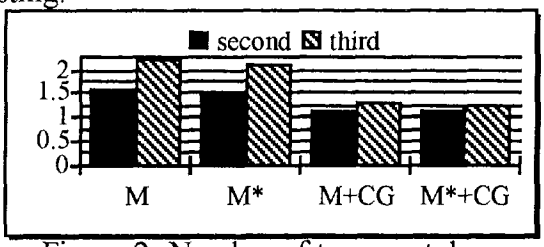

Figure 2- Number of tags per token.

The initial ambiguity of the training corpus is relatively high, as shown in *fig. 1*, and the average number of tags per token is also higher than in other languages –see *fig. 2*. The number of ambiguity classes is also high –290 and 1138 respectively– and some of the classes in the test corpus aren't in the training corpus, specially in the 3rd level tagset. This means that the training corpus doesn't cover all the phenomena of the language, so we would need a larger corpus to assure that it is general and representative of the language.

We tried both supervised and unsupervised[4] training using the 2nd level tagset and only supervised training using the third level tagset. The results are shown in *fig. 3(S)*. Accuracy is below 90% and 75% respectively. Using unknown words to enrich the lexicon, the results are improved –see *fig. 3(S*)*–, but are still far from the accuracy of other systems.

We have also written some biases –to be exact 11– to correct the most evident errors in the 2nd level. We didn't write more biases for the following reasons:

- They can use just the previous tag to change the probabilities, and in some cases we need a wider context to the left and/or to the right.
- They can't use the lemma or the word.
- From the beginning of this research, our intention was to combine this method with Constraint Grammar.

Using these biases, the error rate decreases by 5% in supervised training and by 7% in unsupervised one –*fig. 3(S+B)*.

We also used biases[5] with the enriched lexicon and the accuracy increases by less than 2% in both experiments –*fig. 3(S+B*)*. This is not a great improvement when trying to decrease an error rate greater than 10%, but the enrichment of the lexicon may be a good way to improve the system.

The logical conclusions of these experiments are:

- the statistical approach might not be a good approach for agglutinative and free-order languages –as pointed out by Oflazer and Kuruöz (1994).
- writing good disambiguation rules may really improve the accuracy of the disambiguation task.

As we mentioned above, it is difficult to define accurate rules using stochastic models, so we use the Constraint Grammar for Basque[6] (Aduriz *et al.*, 1997) for this purpose.

The morphological disambiguator uses around 800 constraint rules that discard illegitimate analyses on the basis of local or global context

---

[3] These measures are taken after the process denoted in each column: *M* → morphological analysis; *M** → morphological analysis with enriched lexicon; *CG* → Contraint Grammar.

[4] Even if we used the same corpus for both training methods to compare the results, the latter performed better using a larger corpus.

[5] These biases were written taking into account the errors made in the first experiment.

[6] The rules were designed having syntactic analysis as the main goal.

conditions. The application of CG formalism[7] is quite satisfactory, obtaining a recall of 99,8% but there are still 2.16 readings per token. The ambiguity rate after applying CG of Basque drop from 41% to 12% using 2nd level tagset and 64% to 22% using 3rd level tagset –*fig. 2*– and the error rate in terms of the tagsets is approximately 1%.

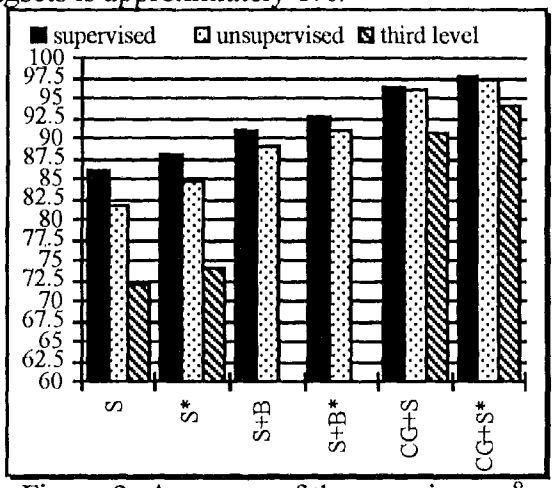

Figure 3- Accuracy of the experiments[8].

5 Combining methods

There have been some approaches to the combination of statistical and linguistic methods applied to POS disambiguation (Leech *et al.*, 1994; Tapanainen and Voutilainen, 1994; Oflazer and Tür, 1997) to improve the accuracy of the systems.

Oflazer and Tür (1997) use simple statistical information and constraint rules. They include a constraint application paradigm to make the disambiguation independent of the rule sequence.

The approach of Tapanainen and Voutilainen (1994) disambiguates the text using XT and ENGCG independently; then the ambiguities remaining in ENGCG are solved using the results of XT.

We propose a similar combination, applying both disambiguation methods one after the other, but training the stochastic tagger on the output of the CG disambiguator.

Since in the output of CG of Basque the average number of possible tags is still high –1.13-1.14 for 2nd level tagset and 1.29-1.3 for 3rd level tagset– and the stochastic tagger produces relatively high error rate –around 15% in 2nd level and almost 30% in 3rd level–, we first apply constraint rules and then train the stochastic tagger on the output of the rule-based disambiguator.

Fig. 1(CG) shows the ambiguity left by Basque CG in terms of the tagsets. Although the ambiguity rate is significantly lower than in previous experiments, the remaining ambiguities are hard to solve even using all the linguistic information available.

We have also experimented with the enriched lexicon and the results are very encouraging, as shown in *fig.3(CG+S*)*. Considering that the number of ambiguity classes is still high –around 240 in the 2nd level and more than 1000 in the 3rd level–, we think that the results are very good.

For the 2nd level tagging, the error rate after combining both methods is less than 3.5%, half of it comes from MORFEUS and Basque CG and the rest is made by the stochastic disambiguation. This is due to the fact that generally the types of ambiguity remaining after CG is applied are hard to solve.

Examining the errors, we find that half of them are made in unknown words trying to distinguish between proper names of persons and places. We use two different tags because it is interesting for some applications and the tagset was defined based on morphological features. This kind of ambiguity is very hard to solve and in some applications this distinction is not important. So in this case the accuracy of the tagger would be 98%.

The accuracy in the third level tagset is around 91% using the combined method, which is not too bad bearing in mind the number of tags –310–, the precision of the input –1.29 tags/token– and that the training corpus does not cover all the phenomena of the language[9]. We want to point out that the experiments with the 3rd level tagset show even clearer that the combined method performs much better than the stochastic. Moreover, we think that CG disambiguation is even convenient at this level because of the initial ambiguity –63%.

---

[7] These results were obtained using the CG-2 parser, which allows grouping the rules in different ordered subgrammars depending on their accuracy. This morphological disam-biguator uses only the first two subgrammars.

[8] *S* '→ stochastic; * → with enriched lexicon; *B* → with biases; *CG* → Constraint Grammar.

[9] In a corpus of around 900,000 words we found 185 new tags and more than 1700 new classes.

Conclusion

We have presented the results of applying different disambiguation methods to an agglutinative and highly inflected language with a relatively free order in sentences.

On one hand, this latter characteristic of Basque makes it difficult to learn appropriate probabilities, particularly first order stochastic models. We solve this problem in part with CG for Basque, which uses a larger context and can tackle the free word-order problem.

However, it is a very hard work to write a full grammar and disambiguate texts completely using CG formalism, so we have complemented this method with a stochastic disambiguation process and the results are quite encouraging.

Comparing the results of Tapanainen and Voutilainen (1994) with ours, we see that they achieve 98.5% recall combining 1.02-1.04 readings from ENGCG and 96% accuracy in XT, while we begin with 1.13-1.14 readings, the quality of our stochastic tagger is less than 90% and our result is better than 96%.

Unlike Tapanainen and Voutilainen (1994), we think that training on the output of the CG the statistical disambiguation works quite better[10], at least using such a small training corpus. In the future we will compile a larger corpus and to decrease the number of readings left by CG.

On the other hand, we think that the information given by the second level tag is not sufficient to decide which of the choices is the correct one, but the training corpus is quite small. However, translating the results of the 3rd level to the 2nd one we obtain around 97% of accuracy. So, we think that improving the 3rd level tagging would improve the 2nd level tagging too. We also want to experiment unsupervised learning in the 3rd level tagging with a large training corpus.

Along with this, the future research will focus on the following processes:

- morphosyntactic treatment for the elaboration of morphological information (nominalisation, ellipsis, etc.).
- treatment of multiword lexical units (MWLU). We are planning to integrate this module to process unambiguous MWLU, to decreases the ambiguity rate and to make the input of the disambiguation more precise.

10 With their method accuracy is 2% lower.

Acknowledgement

We are in debt with the research-team of the General Linguistics Department of the University of Helsinki for giving us permission to use CG Parser. We also want to thank Gilbert Robert for tuning TATOO.

References

Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R. (1996) *EUSLEM: A lemmatiser/tagger for Basque.* EURALEX.

Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M. (1995) *Different issues in the design of a lemmatizer/tagger for Basque.* "From text to tag" SIGDAT, EACL Workshop.

Aduriz, I., Arriola, J.M., Artola, X., Díaz de Illarraza, A., Gojenola, K., Maritxalar, M. (1997) *Morphosyntactic Disambiguation for Basque based on the Constraint Grammar Formalism.* RANLP, Bulgaria.

Alegria, I., Sarasola, K., Urkia, M. (1996) *Automatic morphological analysis of Basque.* Literary and Linguistic Computing Vol 11, N. 4.

Armstrong S., Russel G., Petitpierre D., Robert G. (1995) *An open architecture for Multilingual Text Processing.* EACL'95. vol 1, 101-106.

Brill E. (1992) *A simple rule-based part of speech tagger.* ANLP, 152-155.

Church K. W. (1988) *A stochastic parts program and phrase parser for unrestricted text.* ANLP, 136-143.

Cutting D., Kupiec J., Pedersen J., Sibun P. (1992) *A practical part-of-speech tagger.* ANLP, 133-140.

Karlsson F., Voutilainen A., Heikkilä J., Anttila A. (1995) Constraint Grammar: Language-independent System for Parsing Unrestricted Text. Mouton de Gruyter.

Leech G., Garside R., Bryan M. (1994) *CLAWS4: The tagging of the British National Corpus.* COLING, 622-628.

Oflazer K., Kuruöz I. (1994) *Tagging and Morphological Disambiguation of Turkish Text.* ANLP, 144-149.

Oflazer K., Tür G. (1997) *Morphological Disambiguation by Voting Constraints.* ACL-EACL, 222-229.

Tapanainen P., Voutilainen A. (1994) *Tagging Accurately - Don't guess if you know.* ANLP, 47-52.