

Disambiguation of case suffixes in Basque

Victor Lascurain (1), Eneko Agirre (1), Mikel Lersundi (1)
Luboš Popelínský (2)

(1) University of the Basque Country, Donostia, Spain

(2) KD Group & NLP Lab

Faculty of Informatics, Masaryk University in Brno, Czech Republic

popel@fi.muni.cz, <http://www.fi.muni.cz/kd>

Le but de ce projet était la classification automatique des cas grammaticaux dans la langue Basque. Pour réaliser ça on a appliqué l'apprentissage inductive (les systèmes Tilde et Timbl). On emploie WordNet pour retrouver les synsets et les hyperonyma des mots dans un contexte. L'exactitude était plus haut que 70% pour Tilde et 63% en cas du Timbl.

Basque

agglutinative language

case suffixes are more or less equivalent to prepositions in other languages

case suffixes also used to mark subject and objects of verbs

here : disambiguation of instrumental

Table 1: Possible interpretations of the instrumental case-suffix (-z).

	Basque	English
theme	Matematikaz asko daki	He's an expert in maths
during-time	Gauetz egin dut	I did it by night
instrument	Euskaraz hitz egin	To speak in Basque
manner	Ahots ozen batez	In a loud voice
cause	Haren aitzakiez nekatuta nago	Sick of his excuses
containing	Edalontzia ardoz beteta dago	The glass is full of wine
matter	Armairua egurrez egina dago	The wardrobe is made of wood

Goal : to find the correct interpretation; semantic disambiguation task

Overview

1. Method
2. Data
3. Learning with Tilde
4. Learning with Timble
5. Conclusion

Method

classify each occurrence of the case suffix into one of possible interpretations
taken into account the context

words in the context annotated ambiguously with the morphological
analyser

Basque WordNet – BasWN – employed for obtaining semantic information

	synset	No. of senses	sens/syn	Words	sens/word
Nouns	27649	48214	1.74	22146	2,17
Verbs	3240	9295	2,86	3155	2,95

Table 2: BasWN stats 2002-01-11

Data

142 correctly classified examples extracted from the monolingual Basque dictionary (Sarasola 1986)

Example: "Bazkaz hornitu." (literary "grass feed")

"<Bazkaz>"

"bazka" IZE ARR DEK INS MG

"bazka" IZE ARR DEK INS NUMS MUGM

"<hornitu>"

"hornitu" ADI SIN AMM PART ASP BURU NOTDEK

"hornitu" ADI SIN AMM PART DEK ABS MG

"hornitu" ADI SIN AMM PART NOTDEK

"<\$.>\$"

PUNT_PUNT

Format for Tilde

```
begin(model(example1)). theme.  
leftCtx([]).  
rightCtx([word(hornitu,  
              [[hornitu,adi,sin,amm,part,asp,buru,notdek],  
               [hornitu,adi,sin,amm,part,dek,abs,mg],  
               [hornitu,adi,sin,amm,part,notdek]])]).  
position(word(bazkaz,  
           [[bazka,ize,arr,dek,ins,mg,aorg,has_mai,def_hasi,notgelgen],  
            [bazka,ize,arr,dek,ins,nums,mugm,aorg,has_mai,def_hasi,  
             notgelgen]])).  
end(model(example1)).
```

Learning with Tilde

Inductive logic programming (ILP) – learns first order logic descriptions from a set of examples and a given background knowledge

Tilde – the ILP system which learns first order logic decision trees (Blockeel & De Raedt 1997)

Background knowledge Accuracy

exists/1, forall/1	47%
exists/2, forall/2	55%
hasSynset/1, hasHypoeronym/1 till 3rd level up	
+ exists/1, forall/1	56%
+ exists/2, forall/2	59%
removing implicit class	71.1% (recall 68.3%)

Table 3: Results for exists/2, forall/2 and the refined WordNet predicates

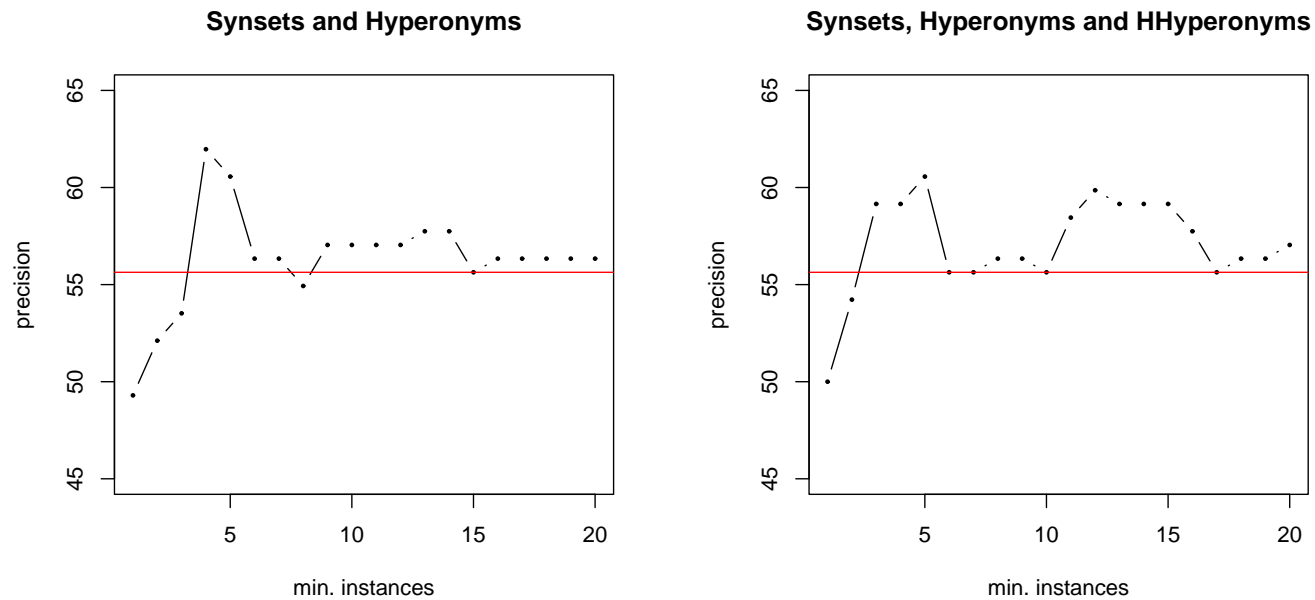
REAL / PRED	cause	containing	instrument	manner	matter	theme	time	total
cause	0	0	1	1	0	3	0	5
containing	0	21	2	0	0	0	0	23
instrument	0	3	7	15	1	3	1	30
manner	0	1	5	29	1	5	0	41
matter	0	3	1	1	1	1	0	7
theme	1	1	1	4	0	23	0	30
time	0	0	0	0	0	1	5	6
total	1	29	17	50	3	36	6	142

Learning with Timble

Timbl (Zavřel & Daelmans 1998) – a program implementing several instance-based learning techniques

It stores the training set in memory, and classifies new cases by extrapolation from the most similar stored cases.

Figure 1: Results for Timbl



synsets/hyperonyms which are true for less than N examples are removed, for $N \in \{1..20\}$.

the horizontal line in the middle shows accuracy 55.6%, the case when no WordNet information has been exploited

When we added the synset information to the data we also add a lot of noise. By adding all the synsets of a word the only thing we do is adding all the possible semantic interpretations of a given word. When we restrict the minimum number of examples in which a synset must be present (the N parameter) data become less ambiguous. Those synsets which belong to different words have better chances to survive. The accuracy increases until we begin to destroy more information than noise. As N moves from 1 to 20 there is a balance between noise and information. The first peak could be due to the situation in which the synset and hyperonym information weigh more than the ambiguity they introduced. From that moment we begin to destroy information

Conclusion

disambiguation of cases in Basque, i.e. automatic classification of case-suffixes, by means of learning techniques and using semantic information from WordNet

47–55% using simple morphological predicates

59% after adding WordNet predicates

71.1% (recall 68.3%) after removing the implicit class

both morphological and WordNet predicates are useful

Future work

use of data from a Basque corpus

better exploitation of WordNet information and/or use of a better ontology