

A 3-step algorithm for morphological disambiguation using untagged corpora

Anna PAPPA
Group C.S.A.R. – Dept of Computer Science
University of Paris 8, France
Tel. : +33149406412 Fax : +33149406783
ap@ai.univ-paris8.fr

Abstract

This article presents a three steps algorithm for morphological disambiguation between the definite article and the personal pronoun in French language. Tested accuracy in a large untagged corpora exceeds 98% with less than 1% of error. Our method has been also experimented on unlabeled Greek corpora and the results prove the system's portability to other languages with similar structure. Not any prior knowledge is available. The rule-based procedure is robust and self-correcting. It can also be used as a shallow parser for verbal and nominal groups identification. The last step of the algorithm consists on the creation of a dictionary with classification of the entries in two grammatical categories : nominal and verbal.

Keywords

Natural Language Processing (NLP), computational morphology, disambiguation rules, untagged corpora, distributional analysis.

1. Introduction

Word-forms are often ambiguous. The sentential context normally decides which analysis is appropriate, this is called disambiguation. And that is one of the problems that a conventional parser has to deal with before tagging. This paper presents an algorithm that can accurately disambiguate with a success rate of 98% the homonymy between the definite article and the personal pronoun, (see table 1) found into both French and Greek languages. This case of ambiguity shows that the same phonetic reality corresponds in two radically different

significations. Here for both languages the definite article is a determinant and the personal pronoun is a referent, it replaces a nominal or verbal group used in the context before.

The method uses the technique of alignment and avoids the need for costly hand-tagged training data, using a small lexicon containing only the grammatical words. Languages offer regularities and thus the possibility of generating rules that can resolve the ambiguity. The procedure is based on the statistics issued from the distributional analysis [1] of the grammatical words in different context.

Table 1. Articles or pronouns in French and Greek.

	French	Greek
singular	le, la, l'	t??, t??, t?, t??, t??
plural	les	t???, t??, ta

There are two basic methods for morphological disambiguation : rule-based [2] and probabilistic or stochastic methods [3]. Mostly of the disambiguation rules taggers are inspired by the Constraint Grammar [4]. Recently a Systemic Grammar [5] proposed a compositional modelling technique tested on grammatically tagged corpora. Even when the disambiguation rules are automatically acquired [6], they are based on tagged text. The probabilistic ones are dominant since 1980s, one of the earliest is the system CLAWS [7], which uses statistical optimisation over n-gram probabilities to assign each word with a tag. An improved version of CLAWS (success rate is 96%) is

used for the 100-million word British National Corpus [8], similar success have been reported for English from others [9]. Most of the stochastic systems derive the probabilities from a hand-tagged training corpus. Some of them based on a Hidden Markov Model can be trained on an untagged corpus with a reported success rate of 96% for English [10], [11], [12]. We also find in [13] the use of the [14] algorithm which disambiguates several types of attachment ambiguities.

2. The algorithm

The system we developed is divided in a three levels analysis (figure 1 shows the algorithm's concept). The first one consists on the execution of the disambiguation rules based on syntactic information using statistical techniques [15]. Not any prior linguistic knowledge [16] is necessary. The second one uses the results provided by the first analysis with rule application priority 1. And the last one searches into the dictionary that the program creates, for any similar word that has already been registered. The training corpus is composed of different kind of texts (but mostly literature and articles) for the French and the

Greek languages.

First we separate the corpus to single phrases. Determining border for sentences is largely developed [17]. All phrases are aligned to the kernel which is the ambiguous word. This technique provides statistical information about grammatical words' context.

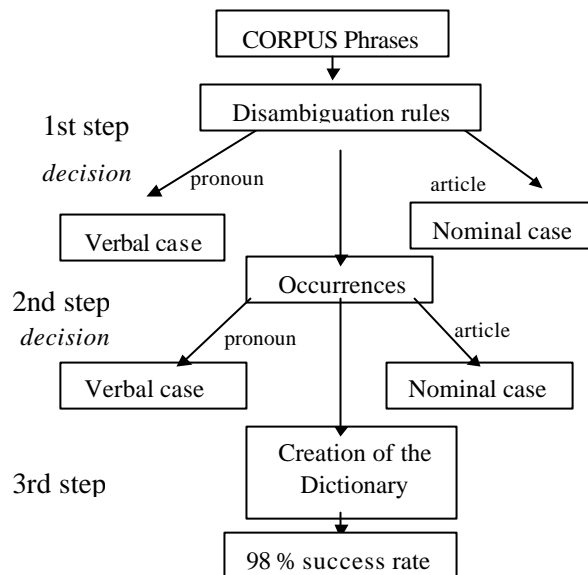


Figure 1. Schema of the algorithm.

Table 2. Extract of the rules.

Rule num	LC3	LC2	LC1	Ambiguous words	RC1	RC2	RC3	Decision	Case	Priority
1	*	*	=Prep	*	*	*	*	article	Preposition + article	1
2	*	*	=Pron	*	*	*	*	pronoun	Pronoun + pronoun	1
3	*	*	=Neg	*	*	*	*	pronoun	Negation + pronoun	1
4	*	*	*	*	=PosPron	*	*	article	Article + Possessive Pronoun	1
5	*	*	=Prep	*	=Suf_Inf	*	*	pronoun	Preposition+pronoun+infinitive	1
6	*	*	=Ger	*	*	*	*	article	Gerund +article nominal	2

2.1.1. 1st step analysis

The first treatment uses the disambiguation rules set (as shown in the table 2 above) to determine the grammatical nature of the ambiguous words. The table has an ergonomical architecture which provides execution rapidity, possibility to add a new rule or to remove an existing one and portability to other languages with similar syntactic structure. The examples shown here are extracted from the French training corpus. The equivalent exists in Greek.

The rules presented above can be applied to both French and Greek corpora. There are about fifty rules for the French and some less for the Greek. The first line of the table defines the context of the cells, LC corresponds to the Left Context of the kernel or ambiguous word (here article or pronoun) and the RC to the Right Context respectively. The alignment of the phrases is showed with the three words that precede and the three that follow the kernel. The annotation with the "=" means that a subroutine treats separately the prepositions, pronouns, negation, suffixes of the infinitive

etc, tests the existence of this type of grammatical words in the phrase and if so the equivalent rule is applied. The “*” means any word of the text. The “decision” column gives the grammatical category of the ambiguous word. The “case” column describes the decision given just before and the last column gives the priority of the rules application. This treatment provides an accuracy of 90,4% of resolved cases. We write “no resolved” in the case cell and 10 in the priority cell to the phrases with the decision cell empty, for further treatment. An extract of the text after the first step analysis is given in the figure 4. Each phrase of the text has a number as the first column shows in the example below.

2.1.2. 2nd step analysis

We have observed that about 10% of the phrases left without any decision, had often similarities, in the RC that immediately follows the kernel (ambiguous word), with the phrases having a decision. Therefore, the second treatment tests only these phrases having an occurrence in the RC1 that has been already treated before. The system applies then the same decision. This step is a self-correcting procedure because if an occurrence found in different phrases has different decisions (either article or pronoun) given by rules with lower priorities, then the decision with priority 1 is the correct one and it is automatically applied to the others phrases where the same word of RC1 occurs. After this treatment the accuracy is significantly grown up to 95,5% of resolved cases.

2.1.2.1. Creation of the dictionary

Before the last step of the analysis we would like to present the idea of the dictionary. We can add new entries to the dictionary at any moment. The procedure is simple. As we already mentioned every rule has a priority tag. Every rule applied the priority tag is given in the flag column. After the first and the second treatment we observe the phrases with their decision after a rule of priority 1 has been applied. The system creates a dictionary with the RC1 entries. That means that the words

found just after the kernel are registered according to their grammatical or verbal or nominal category in a dictionary. An extract is presented in the table 5 below.

Table 5. Extract of the dictionary

num	RC1	Type	Text origin
12	âme	nominal	Luc
62	armée	nominal	Pin-Up
231	bourgeoisie	nominal	Paresse
68	commandement	nominal	Luc
1521	pose	nominal	Furet
312	appliquer	verbal	Chabert
4519	attachait	verbal	Domi
243	pose	verbal	Domi

The first column gives the number of the registered phrase, the RC1 column contains the different lexical entries, the third one describes the grammatical type and the last one gives the text that the phrase comes from.

2.1.3. 3d step analysis

The last step consists on the use of the dictionary. We test if any of the RC1 words figure in the dictionary and if so we apply the decision according to the type of the lexical entry as the table 6 shows below. The system examines only the cases of priority 10 and 5. The 10 corresponds to the “no resolved” cases, but the 5 corresponds to answers given by equivalency, that may means ambiguity in the semantic level which is not the object of the present work, but we would like to use this knowledge for a future work. An example is given below with “pose” which can be either verb or noun. The second step gave the decision pronoun because of the equivalent case found in the number 24. For these cases the priority is always 5. When the third analysis takes place the system looks for “doubles” in both types. If it finds the same entry with two different grammatical decisions it just mentions “ambiguous” in the priority cell, so that we can manually verify the result. This treatment adds a fair 3-4% of solution to our “no resolved” cases and gives the possibility of a future semantic treatment.

3. Conclusion

We have presented an algorithm that can accurately disambiguate the homonymy of the definite article and personal pronoun in French and Greek languages. The system we developed can also be used as a shallow parser to identify nominal and verbal groups [18] with high accuracy (over 95%). After treatment the success rate on completely untagged corpora exceeds 98 % (actually varies between 98 % – 98,5 %) and the error is inferior to 1%. No

prior knowledge is available. The method is easily adapted to other languages having the same structure as we showed for the Greek language. It can also be used for other cases of ambiguity as well as for shallow parsing. The creation of the dictionary at the end of the treatment with the grammatical type of the lemmas gives a further possibility of disambiguation. A future analysis may be based on this for semantical extraction.

Table 3. Extract of the text “le diable amoureux”, after 1st step-analysis.

num	LC3	LC2	LC1	Kernel	RC1	RC2	RC3	Decision	Case	Priority
40	au	fond	et	la	pose	sur	la		No resolved	10
41	la	pose	sur	la	table	assez	près	article	Preposition + article	1
45	finissai t	à	peine	le	commende ment	et	je		No resolved	10
50	moins	pour	savourer	le	tabac	que	pour	article	RC2 relative	2
56	contin uait	en	fumant	la	pipe	que	ma	article	Gerund +article	1
46	je	vois	disparaître	la	pipe	;	et		No resolved	10

Table 4. Extract of the text “le diable amoureux”, after 2nd step-analysis.

num	LC3	LC2	LC1	Ambig word	RC1	RC2	RC3	Decision	Case	Priority
40	au	fond	et	la	pose	sur	la	pronoun	Equivalent n°24 reg	5
41	la	pose	sur	la	table	assez	près	article	Preposition + article	1
45	finissai t	à	peine	le	commendem ent	et	je		No resolved	10
50	moins	pour	savourer	le	tabac	que	pour	article	RC2 relative	2
56	contin uait	en	fumant	la	pipe	que	ma	article	Gerund +article	1
46	je	vois	disparaître	la	pipe	;	et	article	Equivalent n° 56 reg	5

Table 6. Extract of the text “le diable amoureux”, after the 3d step-analysis.

num	LC3	LC2	LC1	Ambig word	RC1	RC2	RC3	Decision	Case	Priority
40	au	fond	et	la	pose	sur	la	pronoun	Equivalent n°24 reg	ambiguous
41	la	pose	sur	la	table	assez	près	article	Preposition + article	1
45	finissai t	à	peine	le	commendem ent	et	je	article	Dictionary n°68	6
50	moins	pour	savourer	le	tabac	que	pour	article	RC2 relative pronoun	2
56	contin uait	en	fumant	la	pipe	que	ma	article	Gerund +article	1
46	je	vois	disparaître	la	pipe	;	et	article	Equivalent n° 56 reg	5

References

- [01] Z. S. Harris, (1954). Distributional Structure, in *Word*, pp. 146-162.
- [02] B. B. Greene and G. M. Rubin, (1971). Automatic grammatical tagging of English. *Technical report*, Brown University.
- [03] E. Charniak, (1997). Statistical parsing with a context free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, AAAI Press / MIT Press, Menlo Park.
- [04] F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila, (1995). Constraint Grammar: A Language-Independent Formalism for Parsing Unrestricted Text. *Mouton de Gruyter*, Berlin, New York.
- [05] D. Hindle, (1989). Acquiring disambiguation rules from text, in *Proceedings ACL 89*, Vancouver Canada pp. 118-125.
- [06] S. Cardey, P. Greenfield, (2003). Disambiguating and Tagging using systemic grammar, in *VIII International Symposium of Social Communication and Applied Linguistics*, actas I pp. 559-564, Santiago de Cuba.
- [07] I. Marshall, (1983). Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus. *Computers in the Humanities*, 17:139—150.
- [08] G. Leech, R. Garside, and M. Bryant, (1994). The large-scale grammatical tagging of text. In N. Oostdijk and P. De Haan, editors, *Corpus-Based Research into Language*, pp. 47—63. Rodopi, Atlanta.
- [09] K. Church (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 136—143, Austin, Texas, ACL.
- [10] J. Kupiec, (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, 6.
- [11] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, (1992). A practical part of speech tagger. In *Proceedings of the 3rd Conference on Applied Language Processing*, pp. 133—140, Trento, Italy.
- [12] D. Elworthy, (1993). Part-of-speech tagging and phrasal tagging. *Technical report*, University of Cambridge Computer Laboratory, Cambridge, England.
- [13] D. Molla, M. Hess, (2000). Dealing with ambiguities in an answer extraction system. In *Workshop on Representation a Treatment of Syntactic Ambiguity in NLP*, pp. 21-24, ATALA, Paris.
- [14] E. Brill and Ph. Resnik, (1994). A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, volume 2, pp. 998-1004, Kyoto, Japan.
- [15] E. J. Briscoe, (1994). Prospects for practical parsing: robust statistical techniques. In P. De Haan and N. Oostdijk, editors, *Corpus-based Research into Language: A Festschrift for Jan Aarts*, pp. 67—95. Rodopi, Amsterdam.
- [16] N. Lindberg and M. Eineborg, (1999). Improving Part of Speech Disambiguation rules by adding linguistic knowledge, in *Proceedings of the 9th International Workshop on Inductive Logic Programming (ILP 99)*. Dzeroski, S. and Flach, P. (Eds). LNAI 1634.
- [17] D. Palmer, M. Hearst, (1994). Adaptive Sentence Boundary Disambiguation, in *Proceedings of the Fourth ACL Conference on Applied NLP*, Stuttgart.
- [18] A. Pappa, (2003). Etiquetage syntaxique automatique des parties du discours en français et en grec. *VIII International Symposium of Social Communication and Applied Linguistics*, actas I pp. 512-517, Santiago de Cuba.