SYNTACTIC APPROACH OF PRONOMINAL ANAPHORA RESOLUTION IN INTELLECTUAL INFORMATION SYSTEMS

Viachaslau Patsepnia

Department of Informatics and Applied Linguistics, Minsk State Linguistic University, 21 Zakharova st., 220 062 Minsk, Belarus, e-mail: *ViacheslavPotsepnia@scnsoft.com*

Abstract. The algorithm, based on syntactic approach to pronominal anaphora resolution in patents in order to resolve anaphoric references and not to lose the information behind those, is described in this paper. The implemented system includes three modules. The first module filters non-anaphoric references, the second one attributes scores to the candidates based on the syntactic information extracted from tagged and parsed text and the third one chooses the candidate which obtained the maximum score. The system was tested on the corpus of US patents.

1. Introduction

Automatic processing of information, represented in natural language both in oral form and in text documents, is one of the most important tasks of computational linguistics. Automatic processing of text documents can be performed for their indexing, summarizing, during machine translation, during knowledge extraction, etc. The common part of these applications is the subsystem of automatic text analysis which is realized at different level of depth of natural language, starting at the morphemic level and ending at the semantic one. The modern information systems become more and more intellectual and require automatic semantic analysis of natural language. Anaphora resolution is a relevant and important problem of semantic analysis since the presence of the unresolved anaphoric references in texts leads to partial loss of information, which, finally, decreases the efficiency indicators of information systems such as recall and precision.

In computational linguistics different approaches to anaphora resolution were elaborated. These approaches can be divided into two groups:

1. traditional algorithms, based on linguistic(syntactical and semantic-syntactical) analysis, discard first impossible candidates and then choose the antecedent on the basis of indicators[1, 2, 4, 5, 6, 9, 10].
2. alternative algorithms (statistical), that choose the most probable antecedent based on the statistical models. [3, 7, 8].

Nevertheless, no famous industrial information system is known where anaphora resolution algorithms were implemented. There is still a need of anaphora resolution systems, which would form part of industrial information system to process correctly information contents avoiding, thus, loss of information beyond anaphoric references.

2. The approach

The suggested approach is directed at the intellectual information systems which perform automatic linguistic analysis of text (that includes part-of-speech tagging and parsing). The approach is purely syntactic unlike the syntactic approach of Lappin & Leass [4], who takes the gender agreement into consideration. The algorithm is directed at the intellectual information systems that work with technical documentation such as patents. The textual contents of this kind of information systems is strictly structured, not subject to metaphors and as patent usually describe method and devices, is unlikely that pronouns “he” or “she”, denoting, as a rule, human beings, will be the central point of a patent, but even “he” or “she” being present, they are likely to be resolved based on syntactic indicators. The suggested algorithm of pronominal (third person personal pronouns are processed only) anaphora resolution takes as input the output of context-dependent parser. The algorithm is implemented in Ruby. The suggested pronominal anaphora resolution system consists of two parts:

a). “it-filter”, which filters pleonastic or non-anaphoric pronoun “it”

b). “resolver”, which analyses the characteristics to the noun phrases and anaphoric pronouns, assigns score based on salience indicators and then chooses the antecedent.

2.1 It-filter

It-filter marks as pleonastic the pronoun “it” if it forms part of the following constructions:

It is **modal_adjective** that

It is **modal_adjective** [for *noun_phrase*] to verb

It is **cognitive_verb_participle**
It seems/appears/requires/follows/implies that
It takes *noun_prhase* to
… makes it **adjective** to *verb*
… makes it *noun_phrase* to *verb*
… finds it **adjective** that
It is *noun_phrase* that
etc.

Modal adjectives are the adjectives that denote modality, e.g. “necessary”, “difficult”, “desirable”, “possible”, etc. Cognitive verb participles are the participles of verbs that describe cognition processes such as “think”, “assume”, “believe”, “suppose”, etc.

2.2 Resolver

The implementation of this part of algorithm was based on metaphor as one of the principle of extreme programming. The resolution process was likened to lawsuit. Resolver works in the context window that is defined in the configuration file. In the experiment the context window embraces 3 sentences including the current one. Resolver includes a set of *detectives*, a set of *juries* and *executor*.

Detectives are the modules which (like human detectives investigate the case) investigate the features of the noun phrases. Two detectives were implemented. ***Anaphoric detective*** investigates the noun phrases and assigns them the feature “anaphoric” or “non-anaphoric”. The feature “anaphoric” is assigned only to the personal pronouns “it” (except those that have been filtered by it-filter), “they”, “them”, “he”, “she”, “him”, “her”. ***Number detective*** investigates the noun phrases and according to the grammatical number assigns them the feature “singular” or “plural”.

After detectives have investigated all possible candidates and attributed features, juries commence their work. Jury (like human juries during an assize decide whether the accused is guilty or not) evaluates the possibility of each candidate to be the antecedent of a certain pronoun taking into consideration one factor or feature. The scores that a jury assigns measure this possibility. The assigned score value may be positive, negative or equal to **nil**. Nil value means that this candidate cannot be the antecedent, i.e. the candidate is filtered. Juries are divided into two classes: unary juries and binary juries (or juries). Unary juries process each noun phrase and do not involve anaphoric pronoun analysis. Resolver includes five unary juries.

I). ***Theme Unary Jury*** gives scores (**+60**) to a noun group in the sentence if it is preceded by subordinating conjunction. E.g. “After *the book* is perforated, it is advanced to the gluing machine.”

II). ***Theme 2 Unary Jury*** gives scores (**+50**) to a noun group which is the subject of a verb phrase with the verb “be” in active voice. E.g. “*The beam splitter* is a complicated apparatus.”

III). ***Theme 3 Unary Jury*** gives scores (**+50**) to a noun group which is the subject of a verb phrase with dialog verbs (“bieleve”, “say”, “agree”). E.g. “IBM Corp. declared it will not accept this pension plan.”

IV). ***Filter Unary Jury*** gives **nil** to impossible candidates such as parenthetic words, figures, pronouns, etc. (e.g. “it”, “he” , “she” , “him”, “her”, “this”, “us”, “one”, “12654”, etc.) E.g. “The value of this variable is *4587*.”

V). ***Repeated Mention Jury*** takes into consideration how many times a noun phrase was repeated in the context window. If the noun phrase was repeated twice or thrice, the jury assigns (14*frequency) scores, if the noun phrase was repeated 4 or more times the jury assigns (7*frequency) scores

Binary juries (or juries) process each pair “candidate-anaphoric pronoun”. Based on the assigned features or other characteristics, the juries analyze both candidate and anaphoric pronoun and assign scores. Resolver includes eight juries:

I. ***Number Jury*** gives scores (**+100**) to the candidates the feature “number” of which coincides with that one of the anaphoric pronoun. In case of incoincidence **nil** is given.

II. ***Distance Jury*** evaluates the distance between the candidate and the anaphoric pronoun. Intrasentential and intersentential distances are considered, as well as the immediate preceding to the anaphoric pronoun:

Table 1. *Distance Jury scores*

Distance factor	**Scores**
same sentence	+80
preceding sentence	0
pre-preceding sentence	-80
the noun phrase preceding the anaphoric noun phrase, or the last noun phrase in the sentence if the anaphoric noun phrase is in a different sentence	+20

III. **CauseEffect Jury** considers the following situations:

Table 2. CauseEffect Jury situations description

subject	object	iobject	subject2	object2	Example
PRESENT	**PRESENT +10**	ABSENT	ABSENT	AnaphoricNP	The PC processes **data** to output *it* in human-readable form later.
PRESENT	ABSENT	**PRESENT +100**	ABSENT	AnaphoricNP	The device is lying on **the table** to make *it* still.
PRESENT	**PRESENT +50**	**PRESENT +50**	ABSENT	AnaphoricNP	The PC outputs **the data** to **the display** to make *it* seen.

IV. **Disjoint Reference Jury** considers the following situations:

Table 3. *Disjoint Reference Jury situations description*

subject	object	iobject	subject2	object2	iobject2	Example
nil	AnaphoricNP					**John** told *him* to get out.
nil		AnaphoricNP				**John** gave the article to *him*.
	nil	AnaphoricNP				John gave the **article** to *him*.
			nil	AnaphoricNP		
			nil		AnaphoricNP	
				nil	AnaphoricNP	John gives the book to Bob to help **his team** in *it*.

V. **Correlation *Jury*** gives scores (**+100**) to a noun phrase if it is followed by a subordinative conjunction (since, while) or an interrogative pronoun (when, how, where) and anaphoric pronoun, e.g. “The machine stamps *the blanks* while they are hot.“

VI. ***Parallel Jury*** gives scores to a noun phrase if it is syntactically parallel to the anaphoric pronoun:

Table 4. *Parallel Jury’s scores*

Syntactic role parallelism	Scores
subject	+80
object	+80
indirect object	+50

VII. ***Possessive Jury*** subtracts score (**-60**) to a noun phrase if it is preceded by possessive pronoun similar to the anaphoric pronoun (it - its; he - him; she - her; they, them - their), e.g. “Car reduces *its weight*, when it slows down.”

VIII. ***Subphrase Jury*** subtracts score (**-80**) to a noun phrases if it is an embedded noun phrase in a sentence that contains anaphoric pronoun in the following constructions: such as …, like…, etc., e.g. “The document such as *a check*, for example, is accepted and then it is returned to the holder.”

All the scores that juries give to the candidate noun phrases were defined experimentally on the corpus.

After juries have processed each candidate and the scores have been assigned, Executor takes the decision which candidate should be chosen as the antecedent. Executor chooses the candidate that obtained the maximum scores on the condition that candidate’s scores overcome the threshold value, which is defined in the configuration file. If two candidates have the same score or the threshold value was not overcome, Executor does not choose any.

3. Evaluation

Since the suggested algorithm is directed at the intellectual information systems that perform linguistic analysis, which includes at least part-of-speech tagging and parsing, and contents of which patents represent, the evaluation was made on the corpus consisting of US patents. The corpus, annotated manually, comprises 20 US patents that contain 959 third person personal pronouns (it, they, them, he, she, her, him). 557 pronouns in the corpus were annotated as non-anaphoric or pleonastic and 402 pronouns in the corpus

were attributed their antecedents (see Tables 5,6,7). Two patents that contain the pronouns “he”, “she”, “him”, “her” were included in the corpus to test the algorithm’s behavior in spite of the fact that it is does not take gender into consideration.

Table 5. *Anaphoric to non-anaphoric pronouns proportion in the corpus*

Type of pronoun	Quantity	Percentage
non-anaphoric	557	49.09%
anaphoric	402	41.91%
Total	**959**	**100%**

Table 6. *Non-anaphoric pronouns in the corpus*

Pronoun	Quantity	Percentage
it	556	99.82%
they	1	0.08%
Total	**557**	**100%**

Table 7. *Anaphoric pronouns in the corpus*

Pronoun	Quantity	Percentage
it	242	60.50%
they	73	18.25%
he/she	46	11.50%
them	33	8.25%
him/ her	8	2%
Total	**402**	**100%**

The performance of It-filter and Resolver were tested separately. Recall and precision were calculated for both of them. The recall for It-filter was calculated as the ratio of correctly filtered non-anaphoric pronouns to all non-anaphoric pronouns in the corpus and the precision was calculated as the ratio of all correctly filtered non-anaphoric pronouns to all filtered pronouns. The recall for Resolver was calculated as the ratio of the number of anaphors resolved correctly to number of all anaphors in the corpus and the precision was calculated as the ratio of the number of anaphors resolved to number of anaphors attempted to be resolved. The total recall and precision for the algorithm were calculated as the simple average of Resolver and It-filter recalls and precisions (see Table 8).

Table 8. *Precision and recall*

Module name	Correctly processed pronouns	Pronouns processed	Recall	Precision
It-filter	549	558	98.56%	98.39%
Resolver	245	397	60.95%	61.71%
Total for the algorithm	**-**	**-**	**79.76%**	**80.05%**

The detailed resolver statistics showed that the recall for “they” constitutes 83.56%, for “them” – 66.67%, for “it” 57.43%, for “him/her” – 50%, for “he/she” – 41.30%.

4. Conclusion

The suggested algorithm is an attempt to improve automatic semantic analysis of the intellectual information systems. The approach represents an attempt to perform semantic analysis on the basis of syntactical information, which can be provided by part-of-speech tagger and parser. The implementation of the algorithm showed that it is a fast, non-expensive and efficient way to resolve pronominal anaphoric references in the technical literature such as patents, which are practically devoid of the pronouns “he/she” and “him/her”, but even if these pronouns are present, as in the described experiment, the algorithm is able to work with the total recall of 79.76% and total precision 80.05%. These indicators are the reliable basis for implementing the developed algorithm into industrial intellectual information systems.

References

[1]. Carbonell, James G. & Ralf D. Brown, “Anaphora resolution: a multi-strategy approach”. *Proceedings of the 12. International Conference on Computational Linguistics(COLING'88)*, Vol.I, (1988), 96-101, Budapest, Hungary,

[2]. Carter, David M., *Interpreting anaphora in naturallanguage texts*. Chichester: Ellis Horwood, (1987)

[3]. Dagan, Ido & Alon Itai, “Automatic processing of large corpora for the resolution of anaphora references”. *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, Vol. III, 1-3, (1990), Helsinki, Finland

[4]. Lappin, Shalom & Herbert Leass, "An algorithm for pronominal anaphora resolution", *Computational Linguistics*, 20(4), (1994), 535-561
[5]. Mitkov, Ruslan, "An integrated model for anaphora resolution", *Proceedings of the15th International Conference on Computational Linguistics (COLING'94)*, (1994), 1170-1176, Kyoto, Japan
[6]. Mitkov, Ruslan, "Anaphora resolution: a combination of linguistic and statistical approaches", *Proceedings of the Discourse Anaphora and Anaphor Resolution (DAARC'96)*, (1996), Lancaster, UK
[7]. Mitkov, Ruslan, "Two engines are better than one: generating more power and confidence in the search for the antecedent", *R. Mitkov, N. Nicolov (Eds), Recent Advances in Natural Language Processing*, John Benjamin Publishers, (1997)
[8]. Nasukawa, Tetsuya, "Robust method of pronoun resolution using full-text information", *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, (1994), 1157-1163, Kyoto, Japan
[9]. Rich, Elaine & Susanne LuperFoy, "An architecture for anaphora resolution", *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-2)*, (1988), 18-24, Texas, U.S.A.
[10]. Rico Pérez, Celia, "Resolución de la anáfora discursiva mediante una estrategia de inspiración vectoral", *Proceedings of the SEPLN'94 conference (SEPLN'94)*, (1994), Córdoba, Spain