Shallow Methods for Named Entity Coreference Resolution

Kalina Bontcheva[1], Marin Dimitrov[2], Diana Maynard[1], Valentin Tablan[1],
Hamish Cunningham[1]
[1]Department of Computer Science, University of Sheffield
211 Portobello St, Sheffield, UK S1 4DP
{kalina,diana,hamish}@dcs.shef.ac.uk
[2]Sirma AI Ltd, Ontotext Lab, 38A Hristo Botev Blvd, Sofia 1000, Bulgaria
marin@sirma.bg

Mots-clefs – Keywords

entités nommées, chaînes de référence, résolveurs d'anaphores
named entities, coreference chains, anaphora resolution

Résumé - Abstract

Nous nous intéressons dans cet article aux méthodes superficielles de résolution d'anaphores et de construction des chaînes de référence, que nous avons développées comme modules du système d'extraction d'information ANNIE. La module "orthomatcher" traite la coréférence orthographique des noms propres et le module de résolution d'anaphores traite les anaphores pronominales dont les antécédents sont des entités nommées. Tous les deux construisent des chaînes de référence, i.e. elles identifient des équivalences entre les entités nommées qui apparaissent dans le texte. Les résultats de précision et de rappel pour l'orthomatcher sont 96/93% pour l'orthomatcher et 66/46% pour la résolution d'anaphores pronominaux.

In this paper we discuss the shallow methods for resolving named entity coreference and building of the coreference chains, which we developed as modules in the ANNIE Information Extraction system. The orthomatcher module deals with detecting orthographic coreference of proper names, while the pronominal resolution module deals with pronominal anaphors, which have named entities as antecedents. Both modules build coreference chains, i.e., identify equivalence between the named entities which appear in the text. The precision/recall results are on average 96/93% for the orthomatcher and 66/46% for the pronominal anaphora resolution.

1 Introduction

Information Extraction (IE) systems are designed to extract fixed types of information from documents in a specific language and domain (Cowie and Lehnert, 1996; Appelt, 1999; Cunningham, 1999). Part of this task is coreference resolution which tries to identify equivalence between named entities that appear in the texts. All references to one and the same entity are grouped into a coreference chain.

In this paper we present the shallow methods for named entity coreference, which we developed as modules in the ANNIE Information Extraction system (Section 2). The orthomatcher module deals with detecting orthographic coreference of proper names (Section 3), while the pronominal resolution module deals with pronominal anaphora, which have named entities as antecedents (Section 4). We also developed a visualisation component where the coreference results can be inspected (Section 5). All these modules can be integrated easily within other systems and applications. The pronominal module can also be extended easily with new rules and heuristics as required. The two modules were evaluated on a corpus which contained newspaper, newswire, and broadcast texts from multiple domains (see Section 6). The results showed that even such shallow and inexpensive methods provide acceptable performance at a minimal processing cost, which is essential when large volumes of data need to be processed.

2 The ANNIE Information Extraction System

ANNIE, A Nearly-New IE system, is provided as part of GATE, a General Architecture for Text Engineering (Cunningham, 2002; Maynard et al., 2002b), which is an architecture, framework and development environment for language processing research and development. Currently ANNIE consists of the following set of modules (which can also be used individually or coupled together with new modules in order to create new applications): tokeniser, sentence splitter, POS tagger, gazetteer, finite state transducer, orthomatcher, and pronominal coreference resolution. The modules communicate via GATE's annotation API, which is a directed graph of arcs bearing arbitrary feature/value data, and nodes rooting this data into document content (in this case text).

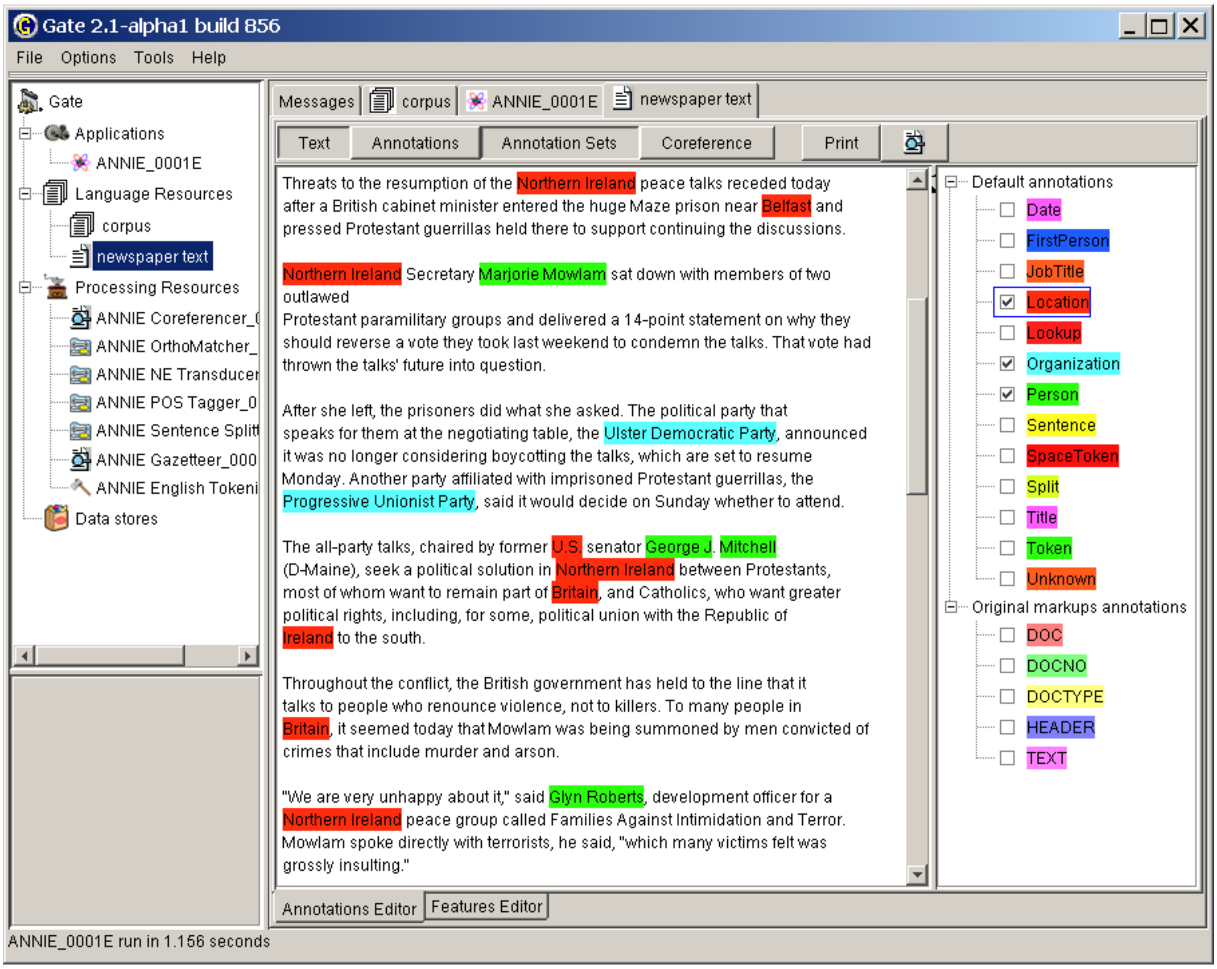
The **tokeniser** splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types (e.g. with an initial capital, all upper case, etc.).

The **sentence splitter** is a cascade of finite-state transducers which segments the text into sentences. This module is required for the tagger. Both the splitter and tagger are domain- and application-independent.

The **tagger** is a modified version of the Brill tagger, which produces a part-of-speech tag as an annotation on each word or symbol.

The **gazetteer** consists of lists such as cities, organisations, days of the week, etc. It not only consists of entities, but also of names of useful *indicators*, such as typical company designators (e.g. 'Ltd.'), titles, etc. The gazetteer lists are compiled into finite state machines, which can match text tokens.

The **transducer** provides finite state transduction over annotations based on regular expressions. For its work the transducer uses grammars that consist of hand-crafted rules written in the JAPE



Threats to the resumption of the Northern Ireland peace talks receded today after a British cabinet minister entered the huge Maze prison near Belfast and pressed Protestant guerrillas held there to support continuing the discussions.

Northern Ireland Secretary Marjorie Mowlam sat down with members of two outlawed Protestant paramilitary groups and delivered a 14-point statement on why they should reverse a vote they took last weekend to condemn the talks. That vote had thrown the talks' future into question.

After she left, the prisoners did what she asked. The political party that speaks for them at the negotiating table, the Ulster Democratic Party, announced it was no longer considering boycotting the talks, which are set to resume Monday. Another party affiliated with imprisoned Protestant guerrillas, the Progressive Unionist Party, said it would decide on Sunday whether to attend.

The all-party talks, chaired by former U.S. senator George J. Mitchell (D-Maine), seek a political solution in Northern Ireland between Protestants, most of whom want to remain part of Britain, and Catholics, who want greater political rights, including, for some, political union with the Republic of Ireland to the south.

Throughout the conflict, the British government has held to the line that it talks to people who renounce violence, not to killers. To many people in Britain, it seemed today that Mowlam was being summoned by men convicted of crimes that include murder and arson.

"We are very unhappy about it," said Glyn Roberts, development officer for a Northern Ireland peace group called Families Against Intimidation and Terror. Mowlam spoke directly with terrorists, he said, "which many victims felt was grossly insulting."

Default annotations
- □ Date
- □ FirstPerson
- □ JobTitle
- ☑ Location
- □ Lookup
- ☑ Organization
- ☑ Person
- □ Sentence
- □ SpaceToken
- □ Split
- □ Title
- □ Token
- □ Unknown

Original markups annotations
- □ DOC
- □ DOCNO
- □ DOCTYPE
- □ HEADER
- □ TEXT

ANNIE_0001E run in 1.156 seconds

Figure 1: Named Entities recognised by ANNIE

(Java Annotations Pattern Engine) language (Cunningham et al., 2002). JAPE is a version of CPSL (Common Pattern Specification Language) (Appelt, 1996) and is used to describe patterns to match and annotations to be created as a result. These grammars identify the named entities in the text, such as organisation, person, location, date (see Figure 1).

The **orthomatcher** and the **pronominal coreference module** are described next.

3 Resolving Proper Name Coreference

The orthomatcher module detects orthographic coreference between named entities in the text, e.g., *James Somebody* and *Mr. Somebody*. Currently it has a set of hand-crafted rules, some of which apply for all types of entities, while others apply only for specific types, such as persons or organisations. The majority of these rules were originally developed in the LaSIE system (Humphreys et al., 2000), while several new ones were added during our work on the MUSE system, which performs robust multi-genre content extraction (Maynard et al., 2002a). Because the MUSE system runs on longer texts, we discovered some problems with the original LaSIE rules, so we restricted the cases when they applied (see rule description below). Also, previously, the rules were always assumed to be transitive, i.e., if name A matches name B, and name B matches name C, then name A matches name C too, without having to test with rules whether this is the case. However, we discovered that there are rules where transitivity should not be assumed and full matching of all entities in the chain should be checked against the rules. For example, the last token rule (see below), assuming transitivity, would match first *BBC News* with *News*, then match *News* also with *ITV News*, which implies wrongly that *ITV News* matches *BBC News*. So instead, each rule in the orthomatcher now has a transitivity flag and this example can be handled correctly, because before constructing the coreference chain *BBC News, News, ITV News BBC News* and *ITV News* are checked for matching, which fails.

We also made the new module Unicode-aware, so it is now capable of handling multiple languages. It also handles case-sensitivity differently: now the orthomatcher can be made to ignore case for all rules, with a centralised parameter, which has led to some reduction in the number of rules needed. Also, in order to improve the module's performance, it was implemented to use the token information produced by the tokeniser module, rather than tokenise the named entity strings internally.

3.1 Orthomatcher Rules

The rules that apply for all types of named entities are:

- *exact match*: two identical occurrences of the same named entity corefer.
- *equivalent*, as defined in a synonym list: this rule is used to handle matching of names like *IBM* and *The Big Blue*. This synonym list can be extended easily with new equivalent names without recompiling the orthomatcher.
- *possessives*: handles named entities in possessive form, e.g., *New York* and *New York's*.
- *spurious*, as defined in a list of spurious names. This rule prevents matching entities which have similar names but are otherwise different. Most frequently this is needed for companies, where a daughter company's name contains the parent company name, e.g., *BT Cellnet* and *BT*.

Some of the rules that apply to organisations and persons are:

- *word token match*: do all word tokens match, ignoring punctuation and word order, e.g., *Kalina Bontcheva* and *Bontcheva, Kalina*.
- *first token match*: does the first token in one name match the first token in the other, e.g., *Peter Smith* and *Peter*. This rule, however, which was originally developed in LaSIE to apply for both people and organisations, had to be modified in order to work correctly for people, because it also matched wrongly *Peter Kline* and *Peter Smith*. This problem was corrected by allowing this rule to fire for persons, only if the shorter name has one token.
- *acronyms* (organisations only): handles acronyms like *International Business Machines* and *IBM*.
- *last token match*: does the last token in one name match the other name (which must be one token only), e.g., *John Smith* and *Smith*.

- *prepositional phrases*: matches organisation names which are inverted around a preposition, e.g., *University of Sheffield*and *Sheffield University*.
- *abbreviations*: matches organisation names, where one name is an abbreviation of the other, e.g., *Pan American* and *Pan Am*.
- *multi-word name matching*: there are several new rules for matching person names when they have more than two tokens and also multi-word company names. The most important rule here is whether all tokens in the shorter name match tokens in the longer name, e.g., *Second Force Recon Company* and *Force Recon Company*.

3.2 Classifying Unknown Proper Names via the Orthomatcher

The orthomatcher can also be used to classify unknown proper names and thereby improve the name recognition process. During the named entity recognition phase, some proper nouns are identified but are simply annotated as Unknown, because it is not clear from the information available whether they should be classified as an entity, and if so, what type of entity they represent. A good example of this is a surname appearing on its own without a title or first name, or any other kind of indicator (such as conjunction with another name, or context such as a jobtitle).

The orthomatcher tries to match Unknown annotations with existing annotations, according to the same rules as before. If a match is found, the annotation type is changed from Unknown to the type of the matching annotation, and any relevant features (such as gender of a Person) are also added to match. Two Unknown annotations cannot be matched with each other. Also, no annotation apart from an Unknown one can be matched with an existing annotation of a different type, e.g. a Person can never be matched with an Organisation, even if the two strings are identical, and its annotation type cannot be changed by the orthomatcher. So, for example, “Smith” occurring on its own in the text might be annotated by the JAPE transducer as Unknown, but if “Mr Smith” is also found in the text (and annotated as a Person), the orthomatcher will find a match between these two strings, and will change the Unknown annotation into a Person one.

4 Resolving Pronominal Coreference

This work falls under the class of ”knowledge poor” approaches to pronominal resolution, which are intended to provide inexpensive (in terms of the cost of development) and fast implementations that do not rely on complex linguistic knowledge, yet they work with sufficient success rate for practical tasks (e.g., (Mitkov, 1998)).

Our approach is similar to other salience-based approaches, which perform resolution following the steps:

- identification of the context of the pronoun;
- inspecting the context for candidate antecedents that satisfy a set of consistency restrictions;
- assigning salience values to each antecedent based on a set of rules and factors

- choosing the candidate with the best salience value.

The implementation relies only on the part-of-speech information, named entity recognition and orthographic coreference information. No syntax parsing, focus identification or world-knowledge based approaches, such as e.g., (Humphreys et al., 1998; Lappin and Leass, 1994), were employed.

The module was built to perform pronominal anaphora resolution, using also some JAPE grammar rules to detect quoted speech and pleonastic *it*. These grammars could easily be used outside the pronominal modules, if required. Previous work, such as (Lappin and Leass, 1994), contains patterns about recognising pleonastic *it*, however we found that they are not flexible enough and miss even small variations of the defined patterns (for further detail see (Dimitrov, 2002)).

Detailed corpus analysis revealed that a few simple, salience-based rules could account for the vast majority of pronominal cases. For example, 80-85% of the occurrences of [he,his,she,her] referred to the closest person of the same gender in the same sentence, or, if unavailable, the closest preceding one. In most cases, they referred back to named entities rather than nominal references. Likewise, [it,its] are handled in the same way, but with scope restriction (because there are many nominals). Currently the rules do not allow for cataphora, but occurrences of these were rare in our corpus.

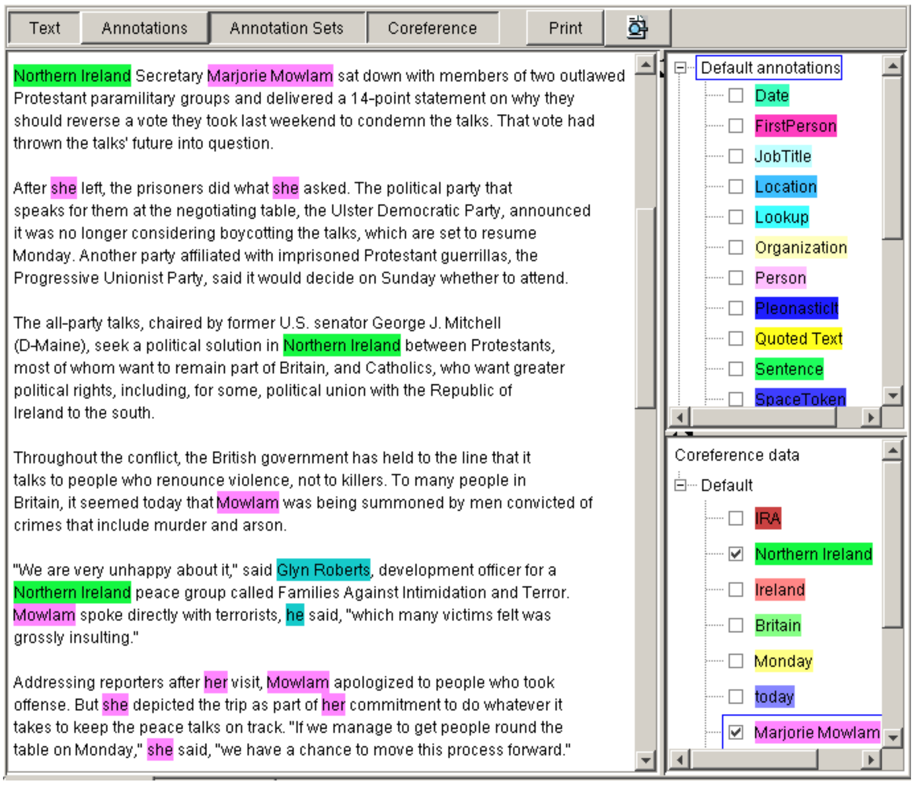
Pronouns occurring in quoted speech are handled by a separate grammar, and require slightly more complex rules. While the context for other pronouns consists of the sentence containing it plus a few of the preceding ones, the context of a pronoun appearing in a quoted fragment is identified according to the quoted span. The context includes only the parts of the sentence containing the pronoun and the one preceding it, which are outside of a quoted fragment. So we consider candidate antecedents (persons in this case) relative to the quoted fragment rather than relative to sentences. So for example, [I, me, my...] generally refer to the pronoun following the quote end in the same sentence or in the one preceding it. If the entire sentence is a quote, i.e. there are no Persons/pronouns that are in the same sentence and out of the quoted fragment, then we look back.

5 Coreference Chain Visualisation

In order to facilitate corpus annotation with coreference data and the debugging process for the coreference modules, we developed a graphical component capable of visualising coreference chains in text. This visual component has been integrated with GATE and becomes available when any of the coreference modules is run (see Figure 2). Each coreference chain is encoded in different colour and can be shown or hidden. When doing corpus annotation, the user can correct wrong chains by selecting them and pressing the Delete key.

6 Evaluation

We evaluated the performance of the orthomatcher by running it on a corpus manually annotated with named entities, and comparing the resulting proper noun coreference chains with



Northern Ireland Secretary Marjorie Mowlam sat down with members of two outlawed Protestant paramilitary groups and delivered a 14-point statement on why they should reverse a vote they took last weekend to condemn the talks. That vote had thrown the talks' future into question.

After she left, the prisoners did what she asked. The political party that speaks for them at the negotiating table, the Ulster Democratic Party, announced it was no longer considering boycotting the talks, which are set to resume Monday. Another party affiliated with imprisoned Protestant guerrillas, the Progressive Unionist Party, said it would decide on Sunday whether to attend.

The all-party talks, chaired by former U.S. senator George J. Mitchell (D-Maine), seek a political solution in Northern Ireland between Protestants, most of whom want to remain part of Britain, and Catholics, who want greater political rights, including, for some, political union with the Republic of Ireland to the south.

Throughout the conflict, the British government has held to the line that it talks to people who renounce violence, not to killers. To many people in Britain, it seemed today that Mowlam was being summoned by men convicted of crimes that include murder and arson.

"We are very unhappy about it," said Glyn Roberts, development officer for a Northern Ireland peace group called Families Against Intimidation and Terror. Mowlam spoke directly with terrorists, he said, "which many victims felt was grossly insulting."

Addressing reporters after her visit, Mowlam apologized to people who took offense. But she depicted the trip as part of her commitment to do whatever it takes to keep the peace talks on track. "If we manage to get people round the table on Monday," she said, "we have a chance to move this process forward."

Figure 2: Coreference Visualisation in GATE

those created by the human annotators. For the evaluation we used the ACE corpus (see (Maynard et al., 2002a)), which consists of newspaper, newswire, and broadcast news. The precision and recall figures are shown in Table 1. The reason for the lower recall was because the ACE corpus considers country names to match their adjectival forms, e.g., *France* and *French*. Consequently, the orthomatcher module was customised for the needs of the project to match such entities by defining them in the equivalence list. This was straightforward and did not require any changes in the orthomatcher code itself. As a result, the recall improved to 96-98%.

Text type	OM	
	precision	recall
broadcast news	94%	92%
newswire	98%	92%
newspapers	98%	95%

Table 1: Precision and recall for the orthomatcher module

The pronoun coreference module was evaluated against a manually annotated part of the ACE corpus, which contained documents from each of the three types. No pronouns were excluded from the evaluation. Occurrences of pronouns not covered by the current implementation degrade the recall. Since nominal antecedents are not considered either, that affects the precision. The results were 66% precision and 46% recall. These numbers are comparable to the performance of other knowledge-poor pronominal coreference implementations (see (Barbu and Mitkov, 2001; Mitkov, 1998)).

The results for each individual group of pronouns are as follows:
]

1. *he, she, his, her, etc.*: 79% precision and 78% recall.
2. *it, its, itself*: 44% precision and 52% recall.
3. *I, me, etc.*: 78% precision and 62% recall.

The results show that the resolution of pronouns from group 1 is relatively successful even with such simple heuristic patterns used and without incorporating any syntactic or semantic information. The precision is degraded by the ratio of nominal antecedents. The algorithm would also benefit from some syntax information indicating the subject of the sentence to be used in combination with recency and gender agreement.

The resolution of pronouns in group 2 is less successful. Apart from the impact of nominal antecedents, additional degradation is induced from errors by the pleonastic *it* module, which although using rules that cover much more cases than the ones in (Lappin and Leass, 1994) still identifies only 38% of the pleonastic occurrences. It is worth noting that the pleonastic *it* module has very high precision and very low recall, so further extension of its patterns will improve the recall and will have a positive impact on the resolution of it.

The very low recall for group 3 is mainly caused by flaws in the quoted speech submodule, which fails to recognize certain quoted constructs. Additionally the performance is negatively impacted by the specifics of the broadcast news texts, where the quoted fragments are not marked, and as a result no attempt for resolution of the pronouns of the 3rd group will be made in this part of the test corpus. Nominal antecedents for pronouns of this group accounted only for 13% of the errors; the rest of the errors were due to incorrectly split sentences and named entities not recognised by earlier ANNIE modules.

If we measure the performance of the pronominal module, independently from the rest of the system, i.e., against the same corpus but with manually annotated named entities, then precision goes up to 73% and recall up to 53% for all pronouns, with the biggest improvement for group 3 where precision goes up to 86% and recall is 76%.

7 Conclusions

The lightweight approach we presented achieves acceptable performance without using any syntax structure information or centering theory methods, which shows that very shallow methods can be sufficient for some coreference tasks. Unfortunately any improvement in the precision and recall just by incorporating further lightweight techniques is unlikely to be achieved. That is why we intend to extend incrementally the basic functionality presented with new features.

In future work, we will address apposition identification, extending the set of handled pronouns, and a nominal coreference resolution module. Appositional coreference has high importance in coreference resolution, because this kind of coreference is observed very often. If syntax information were available for the texts being processed, then apposition could be identified relatively easily. Since ANNIE does not have syntax components, we intend to analyze the corpora in order to find simple heuristic patterns that can identify an acceptable percentage of the appositional occurrences.

In addition, we will extend the set of pronouns being processed. We intend to perform further analysis in order to identify patterns that may help in resolving antecedents for pronouns that are still not handled and that are often observed in text.

Finally, new developments in GATE, such as support for lexical resources like WordNet and also ontologies, will enable us to experiment with a module for nominal coreference based on information from these resources. For example certain candidate antecedents can be identified on the basis of synonymy and hyperonym/hyponymy relations between the words, while ontologies can be useful for resolving certain coreference relations requiring world knowledge.

Références

Appelt, D. (1996). The Common Pattern Specification Language. Technical report, SRI International, Artificial Intelligence Center.

Appelt, D. (1999). An Introduction to Information Extraction. *Artificial Intelligence Communications*, 12(3):161–172.

Barbu, C. and Mitkov, R. (2001). Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of ACL'01*, Tolouse, France.

Cowie, J. and Lehnert, W. (1996). Information Extraction. *Communications of the ACM*, 39(1):80–91.

Cunningham, H. (1999). Information Extraction: a User Guide (revised version). Research Memorandum CS–99–07, Department of Computer Science, University of Sheffield.

Cunningham, H. (2002). GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36:223–254.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., and Ursu, C. (2002). *The GATE User Guide*. `http://gate.ac.uk/`.

Dimitrov, M. (2002). *A Light-weight Approach to Coreference Resolution for Named Entities in Text*. MSc Thesis, University of Sofia, Bulgaria. `http://www.ontotext.com/ie/thesis-m.pdf`.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., and Wilks, Y. (1998). Description of the LaSIE system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. `http://www.itl.nist.gov/iaui/894.02/-related_projects/muc/index.html`.

Humphreys, K., Gaizauskas, R., and Cunningham, H. (2000). LaSIE Technical Specifications. Technical report, Department of Computer Science. University of Sheffield.

Lappin, S. and Leass, H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20:535–561.

Maynard, D., Cunningham, H., Bontcheva, K., and Dimitrov, M. (2002a). Adapting A Robust Multi-Genre NE System for Automatic Content Extraction. In *submitted to AIMSA-02*.

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., and Wilks, Y. (2002b). Architectural elements of language engineering robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*. forthcoming.

Mitkov, R. (1998). Robust Anaphora Resolution with Limited Knowledge. In *Proceedings of COLING'98/ACL'98*.