Named Entity Recognition in Romanian using GATE

Oana Hamza and **Kalina Bontcheva** and **Diana Maynard** and
Valentin Tablan and **Hamish Cunningham**
Dept. of Computer Science, University of Sheffield
Regent Court, 211 Portobello St
Sheffield, S1 4DP, UK
[oana,kalina,diana,valyt,hamish]@dcs.shef.ac.uk

Abstract

This paper presents a Romanian Named Entity recognition system which was developed by reusing and extending IE components developed for English, as part of the MUSE IE system. The system was evaluated on a corpus of diverse text types – religion, news, and fiction. Both the system and the corpus are freely available[1] and were developed using GATE's reusable set of components, its pattern-action rule engine, and the graphical development environment.

1 Introduction

One of the sub-tasks of Information Extraction is the recognition of named entities in texts. Named Entity recognition involves processing a text and identifying certain occurrences of words and expressions as belonging to particular categories of Named Entity (NE), such as locations, persons, organizations, dates, times, monetary amounts and percentages. Named Entity extraction is a key technology in the development of the next generation of information access tools: automatic text summarisation (Maynard *et al.* 02b), information retrieval, etc. For more information about what NE is useful for see (Bontcheva *et al.* 02).

As part of evaluating the portability and reuse across different languages of an English multi-genre named entity recognition system – MUSE (Maynard *et al.* 01) – we developed one of the first Named Entity extraction systems for Romanian.

The system employs a conventional rule-based method in that it divides input Romanian text into words and extracts each Named Entity by referencing gazetteer lists and applying pattern-matching rules. The Romanian Named Entity System (RNES) is evaluated over a corpus of diverse text types, using GATE's IE evaluation tools.

[1]To obtain them contact the second or third author.

2 Specific features for Romanian

There is a set of differences between English and Romanian, especially because Romanian is a more flexible language. For example, the Romanian language accepts typical company designators either before or after the name of the company. It is also possible for them to be both before and after the name of the company, e.g. "SC Zahărul SA"(SA and SC are both company designators), "PAS S.A.", "SC DE CONSTRUCTII ERBASU". Another difference is that the Romanian definite article is not a separate word written before the noun. The definite article appears at the end of the word and it can change the form of it (e.g. the indefinite forms: "universitate"("university"), "minister" ("departament"), "munte" ("mountain") and the definite forms: "universitatea"("the university"), "ministerul"("the department"),"muntele" ("the mountain")). An important observation is that not only common nouns have definite forms, but the proper nouns also have the definite forms, e.g. Olt, Olt*ul*, Apuseni, Apuseni*i*, Cluj, Cluj*ul*.

Another particular thing in Romanian is the genitive and dative case of the proper nouns. If a proper noun is in one of these cases then its indefinite form is changed. Like the definite article, the genitive and dative form of the proper nouns are added at the end of the words, e.g. Bucureşti(indefinite form), Bucureşti*ul*(definite form), Bucureşti*ului*(genitive/dative form). Consequently, the gazetteer lists should contain the definitive and genitive/dative forms.

A particular feature is the case of proper names. The genitive/dative form of feminine proper names is created by changing the end of the words (e.g. "cartea Alin*ei*" ("Alina's book"), "casa Ioan*ei*" ("Ioana's house")). But for male proper nouns, the genitive and dative forms are created by adding a word ("lui") before the noun (e.g. "cartea *lui* Alex" (Alex's book)). Both "lui" and "ei" from the end of the feminine

proper names are called in Romanian "posesive article" and are used to define genitive/dative form. But there are a few exceptions: the feminine proper nouns borrowed from other languages such as "Carmen", "Ingrid" do not change the form by adding "ei" at the end of the word in genitive/dative case. These proper names follow the rule of male proper name, e.g. "cartea lui Carmen"("Carmen's book"), "casa lui Ingrid"("Ingrid's house"). Therefore the gazetteer lists will also contain the genitive/dative form of the feminine proper nouns.

In Romanian the order of the words is also different from English (e.g. "Aeroportul Otopeni" ("Otopeni Airport"), "Banca Comercial Română" ("Romanian Commercial Bank")). The usual order of the words in Romanian is noun followed by an adjective, or a list of adjectives. Naturally, the adjective could appear in front of the noun, but this type of construction could be found in fiction novels or poetry. In current use, the descriptive adjectives such as "mare" ("big" or "great"), "frumos" ("beautiful") could be used in front of nouns, e.g. "*Marea* Adunare Naţională" ("The Great National Assembly").

3 The Romanian Corpus

A very important feature of Romanian is the presence of the diacritics, e.g. "F*ă*g*ă*raş", "B*â*rlad", "Dun*ă*rea", "Lacul Roşu" ("Red Lake"), "Rom*â*nia", "Galaţi" etc. In order not to lose these special letters, when saving the corpus as text files it is necessary to use a GATE component, the GATE Unicode Kit (GUK) (Tablan *et al.* 02). This component helps us to save the Romanian texts with diacritics using the right encoding. The Gate Unicode Kit contains a set of input methods which allow the user to enter text in other languages than the default one. This is done by intercepting the events generated by virtual keyboard. At present it consists of input methods for 17 different languages. GUK also provides a simple Unicode-aware text editor. Besides providing text visualization and editing facilities, the GUK editor performs encoding conversion operations.

The corpus used for the Romanian Named Entity system contains a set of texts found on the Internet. The problem with Romanian texts from the Internet is that most of them are written without diacritics. Thus the available corpus is not very big and varied; it contains 335 files with just over 1 million words. Unlike the English corpus, which contains both spoken and written texts, the Romanian corpus is only written.

The Romanian corpus consists of a collection of texts from different domains such as news, religion and fiction.

3.1 News corpus

The news corpus is composed of a set of newspaper articles published in 2001 or 2002. The size of the corpus is 205 articles with almost 300,000 words. The corpus contains different types of news such as local, national, world and sports news. The articles are collected from the following local newspapers: "Ziarul Personal", "Ecouri Cărăşene", "Amprenta" and "Curierul Zilei", which are all of similar style.

The difficulty of the processing of newspaper articles is the variety of the named entities. Because these articles contain both national and world news, named entities found in these articles could be either specific to Romanian or another language. For example, we can find Romanian cities and person names, but we also can find references to any city or person from any country. Hence it is necessary to keep the gazetteer lists from the MUSE project (e.g. lists of English first names and cities). At the same time these lists are complemented with Romanian spelling variants (e.g. "Londra" (London), "Berna" (Bern), "SUA" (USA)). The articles also contain local news, which can refer to small villages, places from that region. We therefore have specific gazetteer lists for names of villages, places and regions in Romania.

3.2 Fiction corpus

This consists of the novel "1984" written by George Orwell and translated by Mihnea Gafita (Orwell 91). This novel is part of MULTEXT-East project (Multilingual Text Tools and Corpora for Central and Eastern European Languages) [2].

Fiction texts tend to use sets of nicknames such as "Fratele Cel Mare" ("Big Brother"), "Rege" ("King") which are not references to person names in other types of texts. In these texts we can find fictitious festivals (e.g. "Saptamana

[2]MULTEXT-East is a project which ran from '95 to '97 and developed language resources for six languages, see http://nl.ijs.si/*ME*/

Urii" ("Hate Week")), fictitious addresses (e.g. "Blocul Victoria" ("Victory Mansions")) and fictitious places (e.g. "Aerobaza Unu" ("Airstrip One"), "Estasia" ("Eastasia")). We have a switch for fiction texts which turns on the use of specific gazetteer lists for names of fictitious nicknames, places and festivals. We do not use these gazetteer lists all the time because the name entities included in them are not used in colloquial Romanian, and their annotations in other texts apart from fiction texts would involve a lower score.

3.3 Religion corpus

Another collection of texts is "Biblia"("The Holy Bible") [3] translated by Dumitru Cornilescu. In order to keep all the Romanian letters this corpus was saved with UTF-8 encoding using the GATE Unicode Editor and divided into small files.

Religious texts involve sets of names not commonly used elsewhere. For example, it is unlikely that names such as "Picol", "Hazo" would be found in non-religious text as persons, or "Ai" and "Dan" as places. On the contrary "Dan" is also a person and "Ai" is an auxiliary verb in colloquial Romanian. For that reason we have specific lists for names of biblical people and places in order to use them only for religious texts.

4 The IE Technology

The technological part of the system is based on:

1. Architecture and infrastructure from GATE (Cunningham *et al.* 02a)
2. IE approach derived from MUSE (Maynard *et al.* 01)
3. Pattern matching technology from JAPE annotation patterns engine (Cunningham *et al.* 02a)

For further technical details see the GATE User's Guide (Cunningham *et al.* 02b) at `http://gate.ac.uk`.

We have built the NE recognizer for Romanian using three main processing resources: a tokeniser, a gazetteer and a finite state transduction grammar. The system was built and tested using GATE's graphical user interface (Cunningham *et al.* 02a).

[3] "Biblia" was collected from on the web page: http://www.geocities.com/*bibliotecaiogos*/bible.html

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 **gazetteer** consists of lists such as cities, countries, person names, organisations, days of the week, etc. The gazetteer lists are compiled into finite state machines, which can match text tokens.

The **grammar** consists of hand-crafted rules describing patterns to match and annotations to be created as a result. Patterns can be specified by describing a specific text string, or annotations previously attached to tokens (e.g. annotations created by the tokeniser, gazetteer, or document format analysis).

4.1 New Components

We kept the MUSE tokeniser for English, but we made a small change. The difference between RNES and MUSE tokeniser is that the RNES recognises two different words when they are separated by a dash, unlike MUSE system which takes both words as one word. We chose this option because dash separates two completed different words in Romanian.

The structure of the **gazetteer** lists and index file is similar to those used in the MUSE system. Some of the MUSE gazetteer lists are kept, some lists are modified and other lists are deleted. The significant difference is the content of the lists. For example, the list of English person names is replaced by a new list of Romanian person names. For lists such as company designators, we added new elements (e.g. S.R.L, S.A., ACC) which are specific to Romanian companies. In fact, most of the lists are replaced by Romanian names (e.g. company list, city list, region list, and government list). But the names of the gazetteer lists, the major type and minor type from the index file are kept the same where possible. New lists such as monthRoman.lst, which contains roman digits, are added to the gazetteer list because Romanian tends to use another way of writing dates (e.g. 3.XI.1999). In order to process fictitious files, we added special lists which contain festivals, regions and nicknames specific to the novel. Because in Romanian the definite article is added onto the end of the noun, it is not necessary to have a determiner list.

The format of the grammar rules and the types of rules are similar to those used in the MUSE sys-

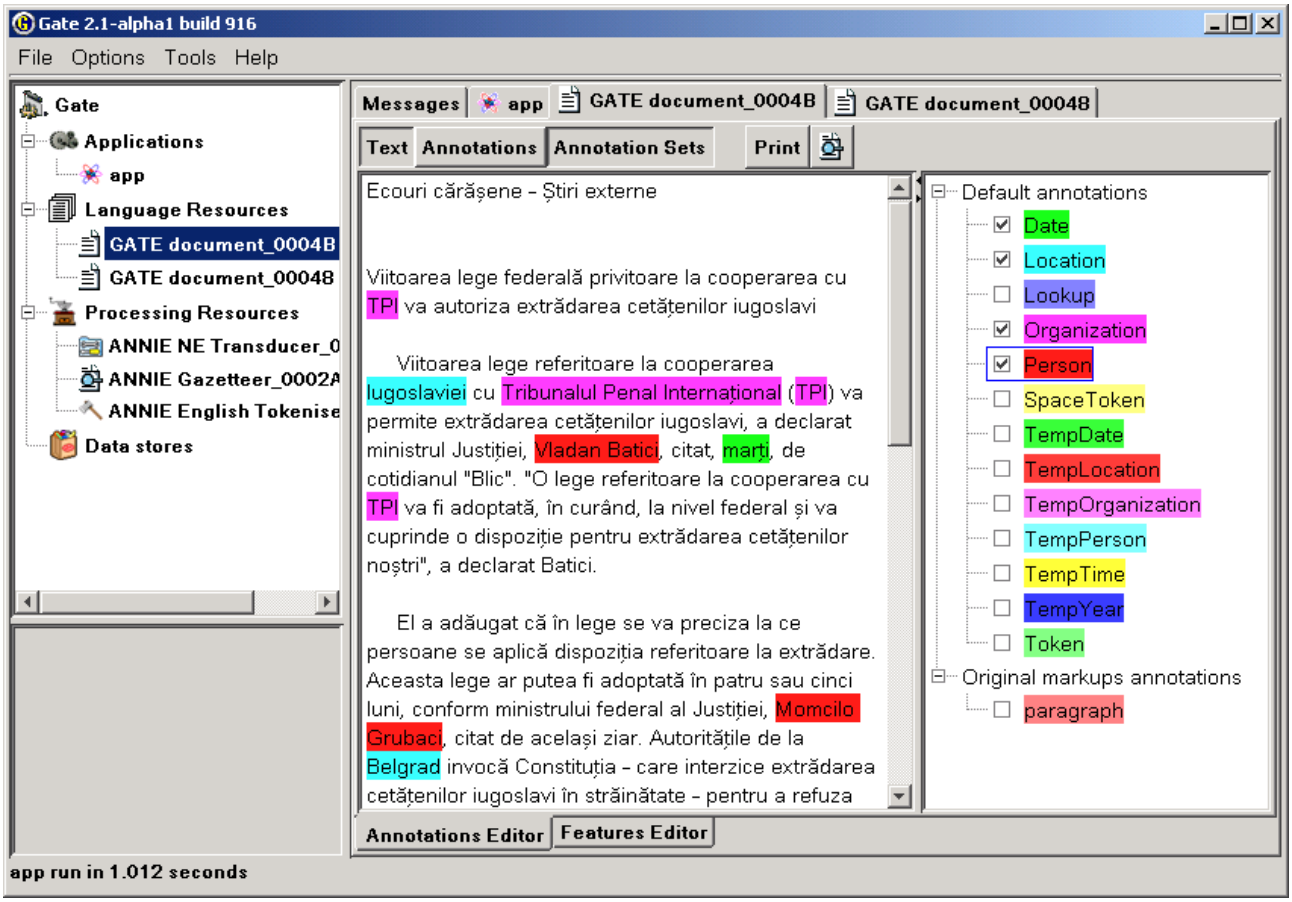

Figure 1: A marked news file in GATE

tem. Additionally, the rules for IP address and Email address from the MUSE system are kept because there is no difference in the way of writing of the IP address and Email address between Romanian and English. Like the gazetteer lists, some of the **grammar** rules are changed; other rules are kept or deleted.

The DateName rule for the determination of date names with the following format "Miercuri 10 Iulie 2000 ("Wed 10 July 2000") is changed because in Romanian between "Miercuri" ("Wed") and the day there could occur a space, comma or dash, but before the year there cannot be a comma like in English. Rules such as ModifierMonth and YearSpan1 are modified because of the presence of the definite article "lui", e.g. "începutul lui Octombrie" ("early October"). In Romanian, the beginning of a month could be expressed using the preposition "de"("of"), e.g. "inceput de Octombrie". Dates such as "10.IX.2000", "miercuri, 14.12.2001" ("Wed, 14.12.2001"), "de vineri pina luni" ("from Friday to Monday"), "08.12 - 14.12.2001", "intre 22 decembrie si 13 ianuarie" ("between 22 December and 13 January") are possible in Romanian; as a result new rules are added to the grammar. New rules are also added for detecting the correct physical addresses: e.g. rules to find the block, floor, and flat number.

The way of writing a phone number in Romanian is quite different from English, therefore most of the phone rules are modified according to the Romanian phone number system.

The possibility of using the context and the priority is also kept for Romanian grammar rules.

For example, the following rule for Organization would mean that a church would be only recognised if it occurs preceded by the words "biserica", "catedrala", "capela" or "Sfântul".

```
MACRO: UPPER\_LETTERS
  ({Token.orth == upperInitial}
   {Token.orth == allCaps\}
   {Token.orth == mixedCaps\})

Macro: SAINT
  ({Token.string == ''St''} ({Token.string == ''.''})?
   {Token.string == ''Sfntul''})
```

```
Macro: CHURCH
  ({Token.string == ‘‘Biserica’’}
   {Token.string == ‘‘biserica’’}
   {Token.string == ‘‘Catedrala’’}
   {Token.string == ‘‘catedrala’’}
   {Token.string == ‘‘Capela’’}
   {Token.string == ‘‘capela’’})

Rule: OrgChurch
Priority: 50
// Biserica Sf\^{a}ntul Ioan
((CHURCH)
 (SAINT)
 (UPPER\_LETTERS)
 (UPPER\_LETTERS)?
):orgName -->
 :orgName.TempOrganization =
   {kind = ‘‘orgName’’, rule = ‘‘OrgChurch’’}
```

Another example of using contextual words is "DayMonthSpan3" rule which recognises a time span such as "între 22 decembrie si 13 ianuarie" ("between 22 December and 13 January").

```
Macro: DAY_MONTH_NUM
  (ONE_DIGIT | TWO_DIGIT)

Macro: MONTH_NAME
  (DATE_PRE
  (({Token.string == ‘‘de’’} |
     {Token.string == ‘‘lui’’})
   )?
  ({Lookup.minorType == month\} |
   {Token.string ==‘‘mai’’} |
   {Token.string ==‘‘Mai’’} |
   {Token.string ==‘‘MAI’’})
   )

Macro: YEAR
  ({Lookup.majorType == year} |
  TWO_DIGIT |
  FOUR_DIGIT |
 ({Token.string == ‘}(TWO_DIGIT)) |
 ({Token.string == ‘‘a’’}DOT
   {Token.string == ‘‘c’’}DOT))

Rule: DayMonthSpan3
// intre 22 dec si 13 ian
(({Token.string == ‘‘ntre’’} |
  {Token.string == ‘‘ntre’’})
  DAY_MONTH_NUM
  MONTH_NAME
  (YEAR)?
  {Token.string == ‘‘i’’}
  DAY_MONTH_NUM
  MONTH_NAME
  (YEAR)?
):date -->
 :date.TempDate = {rule = ‘‘DayMonthSpan3’’}
```

5 Evaluation

A news file annotated by Romanian system can be seen in 1. The system annotated named entities such as person: "Vladan Batici"; organisation: "Tribunalul Penal Internaţional" ("The International Penal Court"); location: "Iugoslaviei" ("Yugoslavia") ; date: "marţi"("Tuesday").

Entity Type	Precision	Recall
Address	0.81	0.81
Date	0.67	0.77
Location	0.88	0.96
Money	0.82	0.47
Organisation	0.75	0.39
Percent	1	0.82
Person	0.68	0.78
Identifier	0.94	0.38
Overall	0.82	0.67

Table 1: Average P + R per entity type, obtained with English NER grammar set

Entity Type	Precision	Recall
Address	0.96	0.93
Date	0.95	0.94
Location	0.92	0.97
Money	0.98	0.92
Organisation	0.95	0.89
Percent	1	0.99
Person	0.88	0.92
Identifier	0.99	0.96
Overall	0.95	0.94

Table 2: Average P + R per entity type, obtained with Romanian NER grammar set

The first experiment in the evaluation of the Romanian Named Entity System is comparing the Romanian grammar against the English grammar on Romanian texts. Firstly, we marked up Named Entities manually in a small corpus which contains online articles from the Romanian newspaper called "Amprenta". Then, we ran over the corpus a system consisting of: the customized Romanian tokenizer, the Romanian gazetteers and the English NE recognition grammar set. By evaluating the results obtained automatically against the human annotations, we were able to get an idea of the performance of the system without modifying the grammar at all (see Table 1).

Then, we run the full Romanian NE recognition system, consisting of the Romanian tokeniser and gazetteers and the Romanian grammar set (see Table 2).

The overall precision and recall scores obtained, when we ran the English grammar over the Romanian text, were quite good, when considered as a first, effortless attempt for performing NE recognition. We have to note that the recall scores obtained were quite low, even in cases of entity types that were identified with great precision

(e.g organisation names); this is due to the fact that many language-specific patterns for identifying the Romanian named entities are not included in the set. Also, patterns that relied on context used the English tokens instead of the Romanian ones, and therefore, the rule conditions were not met. The second table presents the results obtained with Romanian grammars and gives evidence of the very high precision and recall scores that can be obtained rapidly.

We made another test over a few news files containning local news from 2001 from a newspaper called "Ecouri Cărăşene" in order to emphasise the importance of gazetteer lists and to point out small differences between English and Romanian. We ran our system over the news texts using a limited gazetteer, then we ran the system with a full gazetteer. The results are shown in table 3.

The performance on date annotation increased considerably because of new rules for dates of the following type: "9-21 aprilie" ("9-12 April"), "între 5 si 9 noiembrie" ("between 5 and 9 November"), "între 22 decembrie si 13 ianuarie" ("between 22 December and 13 January"). We also added a new rule for "intre Florii si Duminica Tomei" ("between Florii and Duminica Tomei"), where "Florii" and "Duminica Tomei" are two festivals before Easter. The score for person is better because of the improved gazetteer lists (new person names such as "Parris", "Kast", "Gratian"), but it continues to be low because of surnames without context. For organisations and locations, the results are much higher just from completing the gazetteer lists.

We observe that, even with full gazetteer lists the score of person names is quite low because of surnames without contextual clues. In the English system this problem is largely resolved by the use of the Orthomatcher module. For location names table 3 demonstrates the importance of the full gazetteer lists. In this case the recall score increases from 77% to 96%. The organisation names also register a higher score by completing the gazetteer lists. The main reason is that in these texts the organisation names often do not contain special words such as "companie"("company"), "agentia" ("agency"), etc. or company designators such as "SRL", "SA", etc. which provide clues about the entity type.

For testing and development of the system, the corpus was split into three groups: written religious texts (WRIREL), written newspaper texts (WRINEWS), written novel texts (WRINOV). We present below some results for each type of text.

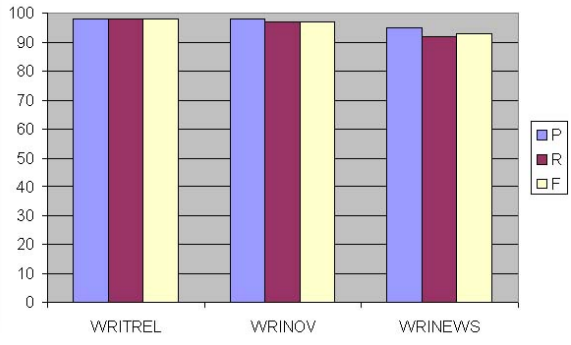

Figure 2: Average results by text type

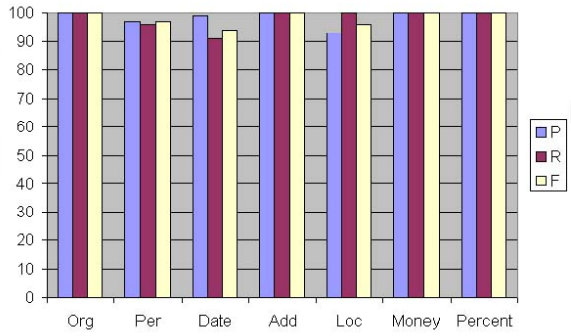

Figure 3: Average results by entity type for WRINOV

Figure 2 shows the average precision, recall and F-measure for all groups of texts - WRINOV, WRIREL and WRINEWS. The highest score was achieved by WRIREL and the lowest score was reached by WRINEWS. This is because the news texts contain all types of named entities, unlike religious texts which contain mostly person names and locations, which were well recognised.

Figure 3 depicts the average (the precision, recall and F-measure) for every entity name from WRINOV group. The system scored well on this type of text, achieving name recognition in the 90s. We have perfect scores for organisations, addresses, money and percent because the number of named entities is very low in this domain. These results are similar to those of the MUSE system (Maynard *et al.* 01),(Maynard *et al.* 02a) or MUC-7 systems, e.g. (Black *et al.* 98).

Annotation Type	Precision		Recall		F-measure	
	Before	After	Before	After	Before	After
Organization	0.93	0.96	0.79	0.90	0.86	0.93
Date	0.90	0.98	0.77	0.98	0.83	0.98
Location	0.96	0.96	0.77	0.96	0.85	0.96
Money	1	1	1	1	1	1
Percent	1	1	1	1	1	1
Person	0.64	0.88	0.65	0.75	0.67	0.81

Table 3: Local news before and after the gazetteer lists were completed and new rules added to the grammar

6 Conclusions and Further Work

This paper has described a system for named entity recognition from Romanian texts. Initial development of this Romanian named entity extraction system has been promising. We identified major problems in each extraction category. The creation of the rules took around 3 weeks; the problem was to create the basic gazetteer lists.

This work was carried out at a time when there was little completed research on Romanian Information Extraction that we could use or compare to. Since then, other researchers have published results on morphological disambiguation for IE in Romanian (part of the SCHUG system (Declerck & Crispi 03)) and date and location identification (Ignat *et al.* 03). Our work could benefit from the integration of SCHUG's morphological component and the large-scale gazetteer lists used by (Ignat *et al.* 03) for locations.

Future development will involve modifying and adding to grammar rules in order to get a higher score. Other avenues for future work include further improvement of the gazetteer lists that the results of the system depend on so heavily. We also plan to extend the corpus to be widely differing in domain, format and genre. This means that we will be able to create specific gazetteer lists and grammar rules for the different genres. GATE's switching controller mechanism will then be used in order to execute the correct set of modules, based on the genre of each text, which is determined using a classifier. This technology has already been used successfully for English within the MUSE system in order to allow specific processing for sport texts and emails.

Acknowledgements

Work on GATE has been funded by EPSRC grants GR/K25267 (GATE), GR/M31699 (GATE2), and GR/N15764/01 (IRC AKT). Support for non-indigenous minority language writing systems within GATE is also funded by the EPSRC on grant GR/N19106 (EMILLE). Several smaller grants supported prototyping work in 1998 and 1999, including two from the US TIPSTER programme and one from the Max Planck Institute in Nijmegen, Holland.

References

(Black *et al.* 98) W. Black, F. Rinaldi, and D. Mowatt. Facile: Description of the named entity system used for muc-7. In *Proceedings of the 7th MUC*, 1998.

(Bontcheva *et al.* 02) K. Bontcheva, D. Maynard, H. Cunningham, and H. Saggion. Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'2002)*, Rome, Italy, 2002.

(Cunningham *et al.* 02a) H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

(Cunningham *et al.* 02b) H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, and C. Ursu. *The GATE User Guide.* `http://gate.ac.uk/`, 2002.

(Declerck & Crispi 03) T. Declerck and C. Crispi. Multilingual Linguistic Modules for IE Systems. In *Proceedings of Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages (IESL'03)*, Borovets, Bulgaria, 2003.

(Ignat *et al.* 03) C. Ignat, B. Pouliquen, A. Ribeiro, and R. Steinberger. Extending and Information Extraction Tool Set to Eastern-European Languages. In *Proceedings of Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages (IESL'03)*, Borovets, Bulgaria, 2003.

(Maynard *et al.* 01) D. Maynard, V. Tablan, C. Ursu, H. Cunningham, and Y. Wilks. Named Entity Recognition from Diverse Text Types. In *Recent Advances in Natural Language Processing 2001 Conference*, pages 257–274, Tzigov Chark, Bulgaria, 2001.

(Maynard *et al.* 02a) D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. Architectural elements of language engineering robustness. *Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data*, 8(2/3):257–274, 2002.

(Maynard *et al.* 02b) Diana Maynard, Kalina Bontcheva, Horacio Saggion, Hamish Cunningham, and Oana Hamza. Using a text engineering framework to build an extendable and portable IE-based summarisation system. In *Proceedings of the ACL Workshop on Text Summarisation*, 2002.

(Orwell 91) George Orwell. *1984*. Univers Publishing House, Bucharest, 1991.

(Tablan *et al.* 02) V. Tablan, C. Ursu, K. Bontcheva, H. Cunningham, D. Maynard, O. Hamza, Tony McEnery, Paul Baker, and Mark Leisher. A Unicode-based Environment for Creation and Use of Language Resources. In *Proceedings of 3rd Language Resources and Evaluation Conference*, 2002.