Multilingual adaptations of ANNIE, a reusable information extraction tool

Diana Maynard, Hamish Cunningham
Dept of Computer Science
University of Sheffield
Sheffield, S1 4DP, UK
diana@dcs.shef.ac.uk

Abstract

In this demo we will present GATE, an architecture and framework for language engineering, and ANNIE, an information extraction system developed within it. We will demonstrate how ANNIE has been adapted to perform NE recognition in different languages, including Indic and Slavonic languages as well as Western European ones, and how the resources can be reused for new applications and languages.

1 Introduction

GATE[1] is an architecture, development environment, and framework for building systems that process human language (Cunningham et al., 2002; Maynard et al., 2002). It has been in development at the University of Sheffield since 1995, and has been used for many R&D projects, including Information Extraction in multiple languages and media, and for multiple tasks and clients. GATE is available freely, as an open source system, under the GNU library licence, and has been downloaded by around 2500 sites worldwide. The core architecture and some applications developed within GATE have been previously demonstrated (Cunningham et al., 2002); however, this demonstration will focus on the multilingual aspects of GATE, and adaptations of its IE system for different languages.

Version 2 of GATE has a large number of added features from the previous version, such as:

- comprehensive multilingual support via Unicode
- tools for performance evaluation
- support for manual annotation
- reusable visualisation components
- database support (Oracle, PostgresQL)
- support for distributed resources from the Web
- comprehensive document format support (SGML, XML, HTML, RDF, email, plain text)

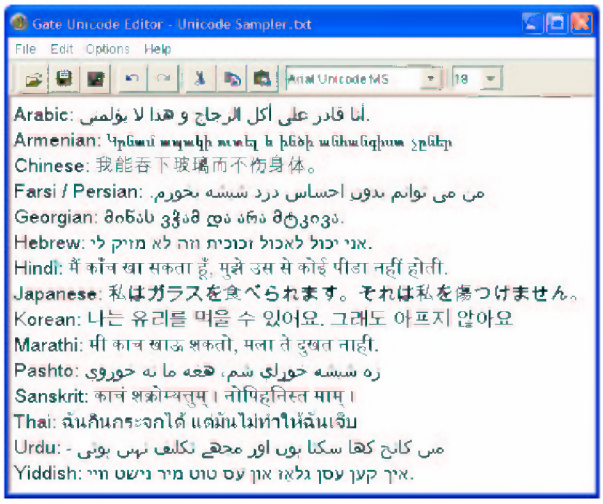

Figure 1: Unicode text in Gate2

2 Processing Resources

GATE provides a baseline set of reusable and extendable language processing components for common NLP tasks, known collectively as ANNIE (A Nearly New Information Extraction System). These include a Unicode tokeniser, sentence splitter, POS tagger, gazetteer, semantic tagger, name coreferencer (orthomatcher)

[1]This work has been supported by the Engineering and Physical Sciences Research Council (EPSRC) under grants GR/K25267 and GR/M31699, and by several smaller grants.

and pronominal coreferencer. For more details, see (Cunningham et al., 2002). ANNIE currently produces precision and recall figures for named entity recognition of around 90%, depending on the text type.

An online demo of ANNIE is available at http://gate.ac.uk/annie/index.jsp. A set of movies demonstrating document and corpus loading, processing and storing, manual annotation of documents and corpora, creating, running, saving and restoring applications and viewing their results is available at http://www.gate.ac.uk/demos/movies.html.

3 Multilingual support - the GATE Unicode Kit

GATE is one of the few architectures to support multilingual processing, using Unicode as its default text encoding. A Unicode enabled graphical user interface (*GUI*) needs to address two main issues: the capability to display text and the ability to enter text in other languages than the default one.

It also provides a means of entering text in a variety of languages and scripts, using virtual keyboards where the language is not supported by the underlying operating platform (Java itself does not support input in many languages covered by Unicode, although it supports Unicode representation). Figure 1 depicts text in various scripts displayed in GATE. The facilities have been developed as part of the EMILLE project (Baker et al., 2002), designed to construct a 63 million word corpus of South Asian languages. There are currently 28 languages supported in GATE, and more are planned for the future. Since GATE is an open architecture, new virtual keyboards can easily be defined by users and added to the system.

Apart from the input methods, GUK also provides a simple Unicode-aware text editor which is important because not all platforms provide one by default or the users may not know which one of the already installed editors is Unicode-aware. Besides providing text visualisation and editing facilities, the GUK editor also performs encoding conversion operations. The editor has proved a useful tool during the development and testing of GATE in a cross-platform environment, while the ability to handle Unicode enables applications developed within GATE to be easily ported to new languages.

4 The future isn't English

Robust tools for multilingual information extraction are becoming increasingly sought after now that we have capabilities for processing texts in different languages and scripts. While the default IE system is English-specific, some of the modules can be reused directly (e.g. the Unicode-based tokeniser can handle Indo-European languages), and/or easily customised for new languages (Pastra et al., 2002). So far, ANNIE has been adapted to do IE in Bulgarian, Romanian, Bengali, Greek, Spanish, Swedish, German, Italian, and French, and we are currently porting it to Arabic, Chinese and Russian, as part of the MUSE project[2].

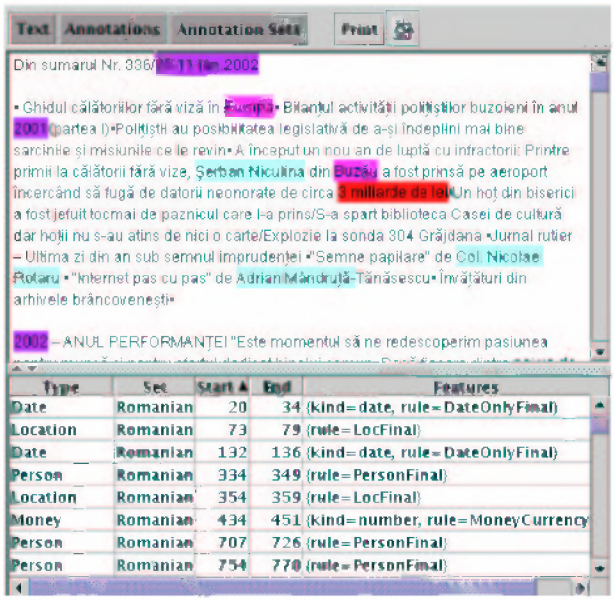

Figure 3: Romanian news text annotated in GATE

4.1 NE in Slavonic languages

The Bulgarian NE recogniser (Paskaleva et al., 2002) was built using three main processing resources: a tokeniser, a gazetteer and a semantic grammar built using JAPE. There was no POS tagger available in Bulgarian, and consequently we had no need of a sentence splitter either. The main changes to the system were in terms

[2]http://www.dcs.shef.ac.uk/research/groups/nlp/muse/

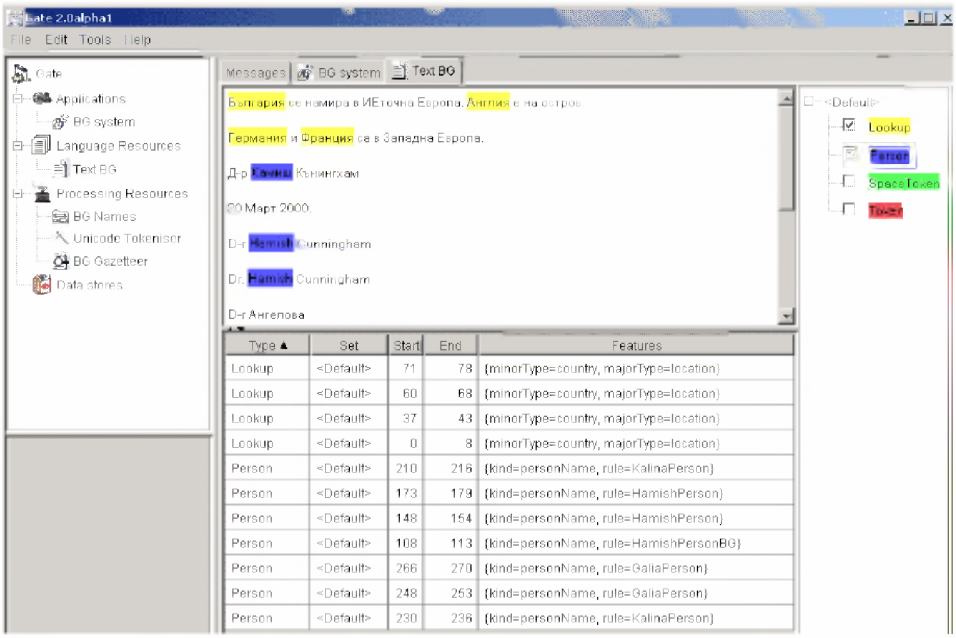

Figure 2: Bulgarian named entities in GATE

of the gazetteer lists (e.g. lists of first names, days of the week, locations etc. were tailored for Bulgarian), and in terms of some of the pattern matching rules in the grammar. For example, Bulgarian makes far more use of morphology than English does, e.g. 91% of Bulgarian surnames could be directly recognised using morphological information. The lack of a POS tagger meant that many rules had to be specified in terms of orthographic features rather than parts of speech. Figure 2 shows some Bulgarian text annotated in GATE.

Since the structure of the Bulgarian and Russian languages is quite similar, we anticipate that converting the Bulgarian system to Russian will be fairly straightforward, and will involve mostly replacing and/or updating gazetteer lists – at least to obtain comparable results.

4.2 NE in Romanian

The Romanian NE recogniser (Hamza et al., 2002) was developed from ANNIE in a similar way to the Bulgarian one, using a tokeniser, gazetteer and a JAPE semantic grammar. Figure 3 shows some Romanian text annotated in GATE.

Romanian is a more flexible language than English in terms of word order; it is also agglutinative e.g. definite articles attach to nouns, making a definite and indefinite form of both common and proper nouns.

As with Bulgarian, the tokeniser did not need to be modified, while the gazetteer lists and grammar rules needed some changes, most of which were fairly minor. For both Bulgarian and Romanian, the modifications necessary were easily implemented by a native speaker who did not require any other specialist skills beyond a basic grasp of the JAPE language and the GATE architecture. No Java skills or other programming knowledge was necessary. The Gate Unicode kit was invaluable in enabling the preservation of the diacritics in Romanian, by saving them with UTF-8 encoding.

4.3 NE in other languages

ANNIE has also been adapted to perform NE recognition on English, French and German dialogues in the AMITIES project[3], a screenshot of which is shown in Figure 4. Since French and German are more similar to English in many ways than e.g. Slavonic languages, it was very easy to adapt the gazetteers and grammars accordingly.

[3] http://www.dcs.shef.ac.uk/nlp/amities

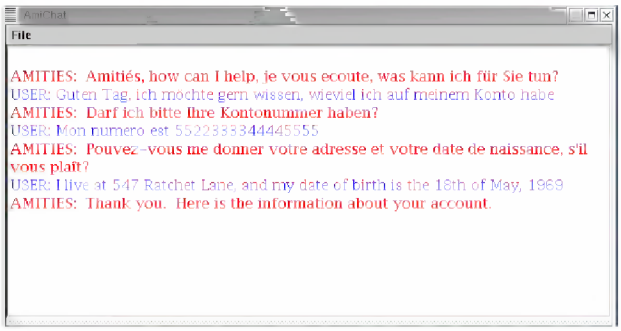

Figure 4: AMITIES multilingual dialogue

4.4 Surprise languages

We are currently investigating methods of adapting ANNIE to new languages with the minimum of resources and time. Our previous experiments with languages other than English have demonstrated that we can get reasonable results in around 2 person months using a native speaker and hand-coded semantic tagging rules, without requiring resources such as dictionaries or POS taggers for that language. We are also planning participation in the TIDES-based "surprise language experiment", which requires various NLP tasks such as IE, IR, summarisation and MT to be carried out in a month on a surprise language, the nature of which will not be known in advance. The open and flexible architecture of GATE, and the separation of linguistic data from processing makes it an ideal environment within which to perform such a task. Any available linguistic resources such as dictionaries and POS taggers can be simply plugged into the model, but if these are not available we can simply modify other components as necessary.

5 Conclusion

In this demo we have shown how an existing set of IE tools has been modified to a diverse set of languages with minimum overhead. The advantage of having such low-overhead portability is that it enables quick deployment of IE tools with acceptable performance, which, even if not developed in end-used applications, can be used to bootstrap the creation of IE-annotated corpora and/or facilitate the training of learning tools for adaptive IE. In addition, some adaptive IE tools are now using the ANNIE components to provide them with richer linguistic information (Ciravegna et al., 2002).

References

P. Baker, A. Hardie, T. McEnery, H. Cunningham, and R. Gaizauskas. 2002. EMILLE, A 67 Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. In *Proceedings of 3rd Language Resources and Evaluation Conference (LREC'2002)*, pages 819–825.

F. Ciravegna, A. Dingli, D. Petrelli, and Y. Wilks. 2002. User-System Cooperation in Document Annotation Based on Information Extraction. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, pages 122–137, Siguenza, Spain.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. 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*.

O. Hamza, D. Maynard V.Tablan, C. Ursu, H. Cunningham, and Y. Wilks. 2002. Named Entity Recognition in Romanian. Technical report, Department of Computer Science, University of Sheffield.

D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. 2002. 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.

E. Paskaleva, G. Angelova, M.Yankova, K. Bontcheva, H. Cunningham, and Y. Wilks. 2002. Slavonic named entities in gate. Technical Report CS-02-01, University of Sheffield.

Katerina Pastra, Diana Maynard, Hamish Cunningham, Oana Hamza, and Yorick Wilks. 2002. How feasible is the reuse of grammars for named entity recognition? In *Proceedings of 3rd Language Resources and Evaluation Conference*.