RC22971 (W0311-036) November 5, 2003
Computer Science

IBM Research Report

The Contribution of Finite-State Technology to Named Entity Recognition and Typing

Branimir K. Boguraev
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

The Contribution of Finite-State Technology to Named Entity Recognition and Typing

Branimir K. Boguraev
IBM T.J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532, USA
E-mail: bran@us.ibm.com

Abstract

This brief note revisits the question of relative merits of manual pattern crafting and machine learning techniques for named entity recognition and typing. In particular, it describes (in outline) an experiment which exemplifies, and empirically validates, the strengths of a combined approach where a robust classification algorithm makes informed use of finite-state grammars defining a number of semantic categories. Assuming the ability to submit a document for analysis by independent devices, one or more of which will be grammar-based, and given a suitable machinery for principled combination of the resulting analysis streams, the experiment demonstrates that high precision pattern-driven semantic category identification (even if the grammars target a subset of the larger set of categories of interest) can significantly boost the overall performance of the combination device.

1 Strategies for Named Entities Recognition

A core component of the KANI[1] project within the NIMD program is an information extraction associate which assumes a capability for identifying text fragments referring to a broad range of ontologically relevant semantic categories. At a broad level of abstraction, this is a generalisation of the Named Entity (NE) identification task, defined and studied originally in the context of the Message Understanding Conference (MUC).

Most of the original research of MUC vintage addressed the named entity recognition problem by pattern matching. Notable examples here include the FASTUS (which pioneered the cascaded model of finite-state processing [Hobbs et al., 1997]), and LTG (which argued for stratification of different rule types, allowing for flexible handling of semantic overlaps [Mikheev et al., 1998]) systems. There are some advantages, and disadvantages, of tackling this problem by purely grammar-based approaches. The early literature focuses at length on rule application strategies, heuristics for resolution of conflicts, and methods for cross-domain adaptation of (usually substantial) rule bases; it is generally agreed that there is some cost to adapting an existing system to new domains.

More recently, efforts have focused on predominantly machine-learning techniques which, given adequate training data, perform at a high-accuracy level (see, for instance, [Bikel et al., 1999], and citations therein). Such levels of accuracy, however, require large amounts of training data; problematically, it is generally hard to quantify, in advance, how large "large enough" is, for any particular application domain, and in correlation with performance levels. For instance, [Ittycheriah et al., 2003] report on a maximum entropy-based approach which, trained over 417 documents, performs at 74.0% F-score. Is this good; how does the score depend on the size of training data; what volume of training data would significantly improve the performance?

Systematic development of rule-based systems, applying the best practices mentioned earlier, demonstrate that if such systems, and their underlying architecture, are well designed, the effort needed for adaptation from a generic NE recognition capability—be it to a new domain, or a different (but related) task—need not be prohibitively expensive; this is the argument made recently by [Maynard et al., 2003], whose MACE system compares favourably to that of [Ittycheriah et al., 2003] above[2].

The purpose of this brief note is not to set the two approaches in opposition. Rather, it starts from the observation that there are still good arguments for using grammar-based named entity recognisers, for a variety of reasons. Typically, these include the fact that certain semantic categories lend themselves well to formal description (consider, for instance, date and time patterns, or monetary amounts), the realisation that certain operational contexts do not

[1] Knowledge Associates for Novel Intelligence.

[2] In the context of the NIST Automatic Content Extraction evaluation, Phase 1; see `http://www.nist.gov/speech/tests/ace/index.htm`. The task, then, is not strictly the MUC NEI task, but is not too dissimilar to it.

always have training data to offer (which may well be the case in situations like rapidly changing information streams arriving at an intelligence analyst's workstation), or the need to distribute a system to external users, who may need to modify and adapt recognisers to their specific data and environments.

The question framed here concerns itself with whether and how to combine the best aspects of manual pattern crafting with machine learning techniques. The 'whether' argument is not particularly new (see, *e.g.* most recently, [Yangarber et al., 2002]). The 'how', on the other hand, raises the issue of optimal combination of inherently different information types. Of particular interest here is a strategy for bringing a continuously updated 'gazetteer' source (*i.e.* an authority file, listing known entities) to the attention of a statistical classifier. Additionally, some conclusions can be made regarding the overall architecture of a combined system, and the way in which independently derived analytical results can be synthesised into a single analysis stream.

2 Combination

A full version of this note will describe an experiment[3] which highlights—and demonstrates, empirically—a particularly cogent reason for developing and deploying grammar-based named entity recognisers. The experiment posited that a named entity recognition device composed of various statistical learning algorithms and pattern-based descriptions could perform better than any of its individual components, by suitably composing and exploiting the individual recognisers' outputs.

In essence, an algorithm developed at IBM Research [Zhang et al., 2003] is capable of examining multiple analysis streams, and learning both from positive and negative examples. Robust Risk Minimization (RRM) is a supervised learning technique, capable of combining diverse information types in a principled way, without concerning itself with the analytical specifics behind individual data streams. Significantly, it makes no assumptions about the distribution of training data. It has been shown to outperform Maximum Entropy [Zhang et al., 2002], and is particularly well suited to the operational constraints of this experiment.

By incorporating a range of grammars—from very high precision rule sets for *e.g.* dates and times, to much looser definitions of other semantic categories, such as roles and places—an RRM-based classifier tunes its parameters to suitable 'trust' thresholds. The experimental setup combines a number of data streams, learning in effect the relative strong (and weak) points of the individual stream generators (*i.e.* NE recognisers) by measuring each stream against the same training data. A model of optimal stream composition is thus derived.

By taking into account the output of (a) rule-based subsystem(s), the framework can fully exploit the advantages of manually-crafted named entity pat-

[3]This is joint work with Nicolas Nicolov and Tong Zhang, of IBM T.J. Watson Research Center.

terns; it also naturally allows for a systematic access to authority file information. This is a point worth emphasising, given that we are separately working on a general, domain-independent method for incrementally updating, and keeping current, gazetteers encoding semantic category type information in rapidly changing domains [Ando, 2003].

3 Finite-State Grammars

The experiment was devised outside of any NIMD-related context. A number of categories (20) were chosen out of a 100-strong inventory of semantic types developed independently, for the purposes of a question-answering system [Prager et al., 2003]; these include, for example, CITY, COUNTRY, STATE, MONEY, ROLE, ORGANISATION, TIME, DATE, DURATION, and so forth.

As mentioned earlier, some of these categories lend themselves quite well to formal description. Additionally, given suitable training data, and a suitable infrastructure for writing annotations-based finite-state (FS) patterns (our finite-state framework is hosted by a document processing environment which has adopted the notion of linguistic annotations as fundamental descriptive / analytic device; see [Neff et al., 2003], [Boguraev and Neff, 2003]), it is feasible to develop recognition grammars with high F-measure as they stand.

Distinctive characteristics of these grammars include cascading (*e.g.* recognising candidate strings prior to commiting to assigning semantic types to them), stratification (where semantic overlaps, and other ambiguities, are largely accounted for by apposite ordering of the recognition grammars), under-specification (coupled with default typing), and heavy dependence on external knowledge sources (in the form of pre-compiled gazetteer files, informed by ontologically mandated properties).

Fourteen of the 20 categories chosen for the experiment were relatively 'close' to some of the categories developed, as a separate project, for NIMD. On the whole, it is not clear how straightforwardly similar semantic labels in different applications / domains map onto each other (indeed, sometimes such a mapping may be far from well defined). In this case, however, it was possible to adapt—with relatively small effort, and using training data independently tagged for the purposes of the experiment—some NIMD grammars so that they were now targeting the 14 category definitions from the QA domain.

The results of the finite-state based analysis are shown below.

```
     AGO: precision: 100.00%; recall:  83.33%; FB1:  90.91
    CITY: precision:  74.89%; recall:  84.46%; FB1:  79.39
 COUNTRY: precision:  92.91%; recall:  83.08%; FB1:  87.72
    DATE: precision:  96.48%; recall:  95.31%; FB1:  95.89
DURATION: precision:  85.10%; recall:  80.82%; FB1:  82.90
   MONEY: precision:  96.94%; recall:  94.32%; FB1:  95.62
NATIONAL: precision:  95.73%; recall:  83.51%; FB1:  89.20
     ORG: precision:  86.00%; recall:  59.34%; FB1:  70.22
```

```
PERSON: precision:  83.21%; recall:  77.50%; FB1:  80.25
  RATE: precision:  97.01%; recall:  95.86%; FB1:  96.43
  ROLE: precision:  84.42%; recall:  84.23%; FB1:  84.32
 STATE: precision:  84.57%; recall:  74.00%; FB1:  78.93
  TIME: precision:  97.06%; recall:  95.65%; FB1:  96.35
  YEAR: precision:  89.13%; recall:  97.04%; FB1:  92.92
```

It is not surprising that for most of the categories the grammars yield higher precison than recall scores. This, however, leads to the observation that this table is broadly indicative of our strategy for content analysis in the NIMD context: the intent is to leverage high-precision grammar-based spotting and recognition for certain 'high-information quotient' items; and then drill in, examining the neighbouring context in more detail.

4 Results from Combination

Perhaps more revealing than the individual scores is the overall result—for all of the 14 categories—of 87.51% precision and 78.52% recall, leading to F-score of 82.77%.

In the setting of this note, however, it is of more interest to note that, as anticipated, the combined NE recogniser has a higher accuracy rate (and F-score) than if trained over 'raw' training data alone. In particular, a configuration utilising a number of input streams *exclusive of* finite-state output is characterised by overall F-score of 82.58%. (This figure is averaged over *all* of the 20 categories.

```
overall      : precision:  83.41%; recall:  81.76%; FB1:  82.58
```

Even if applied to a subset of all the categories—as shown in the previous section—the contribution of FS-based analysis stream to the combination device is significant, as it boosts the combined F-score by 8%:

```
overall       : precision:  91.58%; recall:  89.75%; FB1:  90.66
```

These results demonstrate that a synergistic, hybrid approach to named entity recognition can profitably use parallel data from independently derived streams, gaining substantial improvement in overall performance.

5 Conclusion

A subsequent full paper will present more details of the combination methodology, and of particular strategies for grammar writing, especially adapted for environments where synergistic deployment of (unsupervised and partially-supervised) machine-learning and rule-based techniques are used for named entity extraction of a broad range of semantic categories.

A final point worth mentioning here is that the issues of generating parallel, independently derived yet fully synchronous, analysis streams are not entirely trivial. This is related to the observation made by [Maynard et al., 2003], namely that adaptation of a generic NE recognition capability depends to some extent on the functional granularity of the underlying architecture. We therefore note that in our NIMD work, where the common infrastructure base is provided by a framework for Unstructured Information Management (UIMA) [Ferrucci and Lally, 2003], the right operational assumptions hold for making practical use of a combination device outlined in this note.

References

[Ando, 2003] Ando, R. (2003). E-annotator: semantic similarity among text mentions. Technical report, IBM T.J. Watson Research Center, Yorktown Heights, New York. Internal / unpublished.

[Bikel et al., 1999] Bikel, D., Schwartz, R., and Weischedel, R. (1999). An algorithm that learns what's in a name. *Machine Learning*, 34((1–3)). Special Issue on Natural Language Learning.

[Boguraev and Neff, 2003] Boguraev, B. and Neff, M. (2003). *The Talent 5.1 TFst system: user documentation and grammar writing manual*. IBM T.J. Watson Research Center, Yorktown Heights, New York.

[Ferrucci and Lally, 2003] Ferrucci, D. and Lally, A. (2003). Accelerating corporate research in the development, application and deployment of human language technologies. In *Proceedings of HLT-NAACL Workshop on Software Engineering and Architectures of Language Technology Systems*, Edmonton, ALberta, Canada.

[Hobbs et al., 1997] Hobbs, J., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., and Tyson, M. (1997). FASTUS: a cascaded finite-state transducer for extracting information from natural-language text. In Roche, E. and Schabes, Y., editors, *Finite-state language processing*, Language, Speech, and Communication Series, pages 383–406. MIT Press, Cambridge, MA.

[Ittycheriah et al., 2003] Ittycheriah, A., Lita, L., Kambhatla, N., Nicolov, N., Roukos, S., and Stys, M. (2003). Identifying and tracking entity mentions in a maximum entropy framework. In *HLT-NAACL*, Edmonton, Canada.

[Maynard et al., 2003] Maynard, D., Bontcheva, K., and Cunnigham, H. (2003). Towards a semantic extraction of named entities. In *5th International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, Borovetz, Bulgaria.

[Mikheev et al., 1998] Mikheev, A., Grover, C., and Moens, M. (1998). Description of the LTG used for MUC-7. In *7th Message Understanding Conference (MUC-7)*.

[Neff et al., 2003] Neff, M., Byrd, R., and Boguraev, B. (2003). The Talent system: TEXTRACT architecture and data model. Technical report, IBM T.J. Watson Research Center. Originally presented at HLT-NAACL Workshop on Software Engineering and Architectures of Language Technology Systems, Edmonton, Canada.

[Prager et al., 2003] Prager, J., Chu-Carroll, J., and Czuba, C. (2003). A multi-strategy, multi-quesiton approach to quation answering. Technical report, IBM T.J. Watson Research Center, Yorktown Heights, New York. Submitted for publication.

[Yangarber et al., 2002] Yangarber, R., Lin, W., and Grishman, R. (2002). Unsupervised learning of generalized names. In *19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

[Zhang et al., 2002] Zhang, T., Damerau, F., and Johnson, D. E. (2002). Text chunking based on a generalization of winnow. *Journal of Machine Learning Research*, 2:615–637.

[Zhang et al., 2003] Zhang, T., Damerau, F., and Johnson, D. E. (2003). A robust risk minimization based named entity recognition system. In *CoNLL-2003*.