Entity Extraction is a Boring Solved Problem – or is it?

Marc Vilain
The MITRE Corporation
Burlington Rd
Bedford MA 01730 USA
mbv@mitre.org

Jennifer Su
The MITRE Corporation *and*
Cornell University
Ithaca NY 14853 USA
jfs29@cornell.edu

Suzi Lubar
The MITRE Corporation
Burlington Rd
Bedford MA 01730 USA
slubar@mitre.org

Abstract

This paper presents empirical results that contradict the prevailing opinion that entity extraction is a boring solved problem. In particular, we consider data sets that resemble familiar MUC/ACE data, and report surprisingly poor performance for both commercial and research systems. We then give an error analysis that suggests research challenges for entity extraction that are neither boring nor solved.

1 Background

Entity extraction or named entity recognition, as it is sometimes called, is a known and familiar problem. Named entity (NE) tagging has been the subject of numerous shared-task evaluations, including the seminal MUC 6, MUC 7 and MET evaluations, the CoNLL shared task, the SIGHAN bake-offs, and the ACE evaluations. With this track record, and with commercial vendors now selling named-entity tagging for a fee, many naturally consider entity extraction to be an essentially solved problem. The present paper challenges this view.

The main issue, as we see it, is transfer: NE taggers developed for a specific corpus tend not to perform well on other data sets. Kosseim and Poibeau (2001), for one, show that the informal language of email or speech transcriptions befuddles taggers built for journalistic text. Minkov *et al* (2005) further explore the systematic differences between journalistic and informal texts, training separate taggers for each text source of interest.

Because named entity taggers are so strongly based on surface features, it isn't surprising to observe poor tagger transfer across texts with significantly different styles or with unrelated content. In this paper, we report on the more surprising result that transfer issues arise even for texts with closely aligned content or closely aligned styles.

In particular, we consider a range of primarily business-related texts that are, on the face of it, close in style and/or substance to the journalistic stories in existing NE data sets, MUC 6 in particular. We thus would have expected these texts to support good transfer performance from taggers configured to the MUC task. Instead, we found the same kinds of performance drops as Kosseim and Poibeau had noted for informal texts. Our aim here is to shed light on the how and why of this.

2 Scope of the present study

We begin with a disclaimer. Our goal is not so much to present new technical solutions to NE recognition, as to draw attention to those aspects of the problem that remain unsolved. We cover two main thrusts: (i) a black-box evaluation of several NE taggers (commercial and research systems); and (ii) an error analysis of system performance.

2.1 Evaluation data

Our evaluation data set contains three distinct sections. The largest component consists of publicly-available financial reports filed with the Securities and Exchange Commission (SEC), in particular the 2003 forms 10-K filed by eight Fortune 500 companies. These corporate annual reports share the same subject matter as much business news: sales, profits, acquisitions, business strategies and the like. They take, however, a more technical slant and are rich in accounting jargon. They are also longer, ranging in our study from 22 to 54 pages.

Proceedings of NAACL HLT 2007, Companion Volume, pages 181–184,
Rochester, NY, April 2007. ©2007 Association for Computational Linguistics

Pocahontas	Rule-based
Belle	Rule-based
Jasmine	Statistical, HMM, MUC-trained
Mulan	Statistical, CRF, MUC-trained
Ariel	Rule-based, 10-K tuning

Table 1: system pseudonyms.

Preliminary exploration with our own MUC 6 tagger showed these SEC filings to be particularly hard to tag. Because their sheer length and technical emphasis seemed implicated in this poor performance, we assembled a second corpus of forty Web-hosted business stories from such news providers as MS-NBC, *CNN Money*, and *Motley Fool*. These stories focus on the same eight companies as our 10-K data set, but are shorter and less technical, thus allowing us to isolate length and technicality as factors in tagging business texts.

The final portion of our test set consists of ten news stories that were selected to closely match the kind of data used in past MUC evaluations. They were drawn from the New York Times (NYT) and Wall Street Journal (WSJ) on-line editions, and focus on current events, thus providing one more comparable dimension of evaluation.[1]

2.2 Evaluated systems

Five systems participated in our study, representing a range of commercial tools and research prototypes. Two of these are state-of-the-art hand-built systems based on rule/pattern interpreters. Two are open-source statistical systems, one based on HMMs, and the other on CRFs; both were trained on the MUC 6 data set. The final system is our own legacy MUC-style tagger, noted as *Ariel* in Table 1. Except as noted below, all the systems were run out of the box, with no adaptation to the data.

License and privacy concerns prevent us from identifying all the systems; instead this paper reports most results anonymously, using the names of Disney heroines as system pseudonyms. We have, however exposed the identity of our own system out of fairness, as it benefited somewhat from earlier tuning to SEC forms 10-K.

2.3 Evaluation method

We attempted to replicate the procedure used in the MUC evaluations, extending it only as required by the characteristics of the taggers. The test data were formatted as in MUC 6, and where SGML markup ran afoul of system I/O characteristics, we remapped the data manually, resolving, *e.g.*, crossing tags that may have strayed into the output.

To provide scores that could be compared with the MUC evaluations, we created MUC6-compliant answer keys (Sundheim, 1995), and remapped system output to this standard. We removed system responses that were considered non-taggable in MUC (*e.g.*, URLs) and conflated fine-grained distinctions not made in MUC (*e.g.*, remapping *country* tags to *location*). Scores were assessed with the venerable MUC scorer, which provides partial credit for system responses that match the key in type but not extent, or vice-versa. The scorer also provides a full error analysis, separately characterizing each error in a system response.

3 Findings

Table 2, overleaf, presents our overall findings, aggregated across the three primary entity types: *person*, *organization*, and *location* (the ENAMEX types in the MUC standard). We generally did not measure the MUC TIMEX (*dates*, *times*) and NUMEX types (*moneys*, *percents*) because: (i) neither of the statistical systems generate them; (ii) those systems that do generate them tend to do well; (iii) they are overwhelmingly more frequent in the SEC data than in news, thus skewing results. For completeness' sake, however, Table 2 does provide all-entity news scores in parentheses for those systems that happened to generate the full set of MUC-6 entities.

Turning now to actual performance measurements, Table 2 does not present an especially pretty picture. Aside from two systems' runs on the MUC-like current events, all the scores are substantially below those obtained by competitive MUC systems, which typically reached F scores in the mid-90s, with a high of F=96 at MUC-6.

SEC. The worst performances were turned in for SEC filings, as shown in the first block of rows in Table 2. While precision is generally poor, recall is even worse. One reason for this is the very frequent rightwards shortenings of company names (*e.g.*, from *3M Corporation* to *the Corporation*), in contrast to the leftwards shortening (*e.g., 3M*) favored in news texts. *Ariel* had been tuned to tag all these cases, but the other systems only tagged a scattershot fraction. To isolate the contribution of

[1] We will make the non-copyrighted part of our corpus (the 10-Ks) available to other researchers.

	Pocahontas	Belle	Jasmine	Mulan	Ariel
SEC filings	R=58	R=28	R=50	R=50	R=71
	P=65	P=52	P=43	P=56	P=79
	F=61.1	F=36.4	F=42.7	F=52.6	F=74.5
SEC filings, *"the Corp."* optional	R=71	R=36	R=55	R=60	R=71
	P=65	P=52	P=40	P=56	P=79
	F=68.0	F=42.8	F=46.2	F=57.9	F=74.7
Business news	R=80 (82)	R=64 (69)	R=76	R=65	R=71 (75)
	P=80 (79)	P=86 (83)	P=63	P=74	P=74 (75)
	F=80.1 (81)	F=73.5 (75)	F=69.1	F=69.2	F=72.3 (75)
Current events (MUC-like)	R=94 (94)	R=59 (63)	R=79	R=79	R=89 (91)
	P=94 (93)	P=82 (80)	P=70	P=92	P=91 (92)
	F=94.3 (94)	F=68.5 (71)	F=74.5	F=84.9	F=90.4 (92)

Table 2: aggregated extraction scores, *ENAMEX* only, unless parenthesized (in parens = all entities).

these cases to system recall error, we recalculated the scores by making the cases optional. The scorer removes missing optional responses from the recall denominator, and as expected recall improved; see the second block in Table 2.

Business news. The most consistent performance across systems was achieved with business news, with scores ranging in F=69-80. This is a huge improvement over the gaping F=36-75 range we saw with SEC filings (F=43-75 with optional short names). This confirms that length and financial jargon are implicated in the poor performance on forms 10-K. Nonetheless, these improved scores are still 15-20 points lower than the better MUC scores. Is business language just hard to tag?

MUC-like news. Our attempt to replicate the MUC evaluation data yields an equivocal answer. Two systems (*Pocahontas* and *Ariel*) achieved MUC6-level scores; it may not be coincidental that both are next-generation versions of systems that participated at MUC. Of the other systems, MUC-trained *Mulan* also showed substantial improvement going from business news to current events.

While it is good news that three of the systems that were explicitly trained on MUC (manually or statistically) did well on MUC-like data, it is disquieting to see how poorly this training generalized to other news texts.

4 Factors affecting performance

A finer analysis of our three data sets helps triangulate the factors leading to the systematic performance differences shown in Table 2.

Prevalence of organizations. One factor especially stands out: as Table 3 shows, organizations are twice as prevalent in the business sources as in the MUC-like data. As organization scores generally trail scores for persons and locations (Table 4), this partly explains why business texts are hard.

Kinds of organizations. But that does not explain it all. The profiles in Figure 1 show that current events favor government/quasi-government names (*e.g.,"Congress," "Hamas"*). They are less linguistically productive than the corporate and quasi-corporate names in business texts, and so are more amenable to being explicitly listed in name gazetteers. Florian *et al* (2003) note the effectiveness of gazetteers for tagging the CoNLL corpus.

Editorial standards. Our business news data reflect a growing portion of Web-hosted texts that relax the journalistic editorial rules of traditional news sources such as the NYT or WSJ. For instance, our data show the same frequent omission of corporate designators (*e.g. "inc."*) that Kosseim noted in informal text. Whereas news sources of record will generally mention a company's designator at least once in a story, our business data frequently fail to do so at all, thus removing a key name-tagging cue. By tracing the *Ariel* rule base, we found that the absence of any designator was implicated in 81% of the system's recall error for organization names.

Length. Name taggers often overcome this kind of missing evidence by second-passing a text, propagating name mentions identified in the first

	SEC	Business	MUC
Org	70%	65%	29%
Per	9%	23%	35%
Loc	21%	12%	36%

Table 3: Relative distribution of entity types

	Poca.	Belle	Jasm.	Mul.	Ariel
S org	F=62	F=10	F=46	F=53	F=83
opt	F=74	F=14	F=52	F=61	F=83
S per	F=75	F=65	F=49	F=64	F=60
S loc	F=79	F=77	F=49	F=74	F=78
B org	F=77	F=72	F=70	F=63	F=66
B. per	F=90	F=85	F=70	F=69	F=79
B. loc	F=78	F=76	F=59	F=75	F=73
M org	F=90	F=58	F=48	F=80	F=80
M per	F=99	F=90	F=84	F=81	F=92
M loc	F=98	F=81	F=74	F=89	F=95

Table 4: Type subcores (S=SEC, B=biz., M=MUC)

pass to matching but undetected mentions (Mikheev, 1999). This strategy runs foul, though, when the first pass produces precision errors, as these too can get propagated. Document length is implicated in this through the greater cumulative likelihood of making an error on the first pass and of finding a mention that matches the error on the second pass.

Quasi-names and non-names. A final factor that especially afflicts the Forms 10-K is the similarity of names and non-names. Non-taggable product names (*"AMD Athlon"*) often look like legitimate subsidiaries, while valid operating divisions (*"Health Care"*) are often hard to distinguish from generic designations of market segments.

5 Implications for further research.

What surprised us most in conducting this study was to find so obvious a transfer gap among what appear to be very similar text sources. We were also surprised by the involvement in this of relaxed editorial standards around seeming trivia (like the keyword *"inc."*) This suggests, for one, that current techniques remain too dependent on skin-deep word co-occurrence features. It also suggests that the editorially pristine news texts used in so much NE research may be atypically easy to tag.

While name-tagging programs may struggle with editorially informal texts, the absence of surface contextual cues poses no noticeable challenge to human readers. What cues are left, and there are many, are semantic in nature: predicate-argument structure, selectional restrictions, organization of the lexicon, *etc.* Recent efforts to create common propositional banks and lexical ontologies may thus have much to offer. Indeed, current research in these areas is just beginning to trickle down to the name-tagging problem (Mohit & Hwa, 2005).

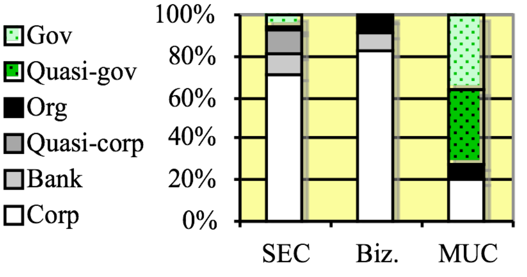

Figure 1. Kinds of organization

Another key issue is ensuring tagging coherency at the whole-document level. This might help alleviate the kind of error propagation with dual-pass strategies that particularly afflicts long documents. Recent applications of statistical co-reference models are beginning to show promise (Finkel *et al*, 2005; Ji & Grishman, 2005).

Lastly, we can see this whole study as a particular challenge case for transfer learning, and indeed such work as Sutton and McCallum's (2005) has looked at the name-tagging task from a transfer learning standpoint.

It may thus be that today's exciting emerging work in "unsolved" areas – semantics, reference, and learning – could come to play a key role in what is sometimes maligned as yesterday's boring solved problem.

References

Finkel J R, Grenager T, Manning C (2005). Incorporating non-local information into information extraction systems by Gibbs sampling, *Proc ACL*, Ann Arbor.

Florian R, Ittycheriah A, Jing H, Zhang T (2003). Named entity recognition through classifier combination, *Proc CoNLL*, Edmonton.

Ji H, Grishman R (2005). Improving name tagging by reference resolution and relation detection, *Proc ACL.*

Kosseim L, Poibeau T (2001). Extraction de noms propres à partir de textes variés. *Proc. TALN,* Toulouse.

Mikheev A, Moens M, Grover C (1999). Named entity recognition without gazetteers. *Proc. EACL*, Bergen.

Minkov E, Wang R, Cohen W (2005). Extracting personal names from email. *Proc HLT/EMNLP*, Vancouver.

Mohit B, Hwa R (2005) Syntax-based semi-supervised named entity tagging. *Proc ACL*, Ann Arbor MI.

Sundheim B (1995), ed. *Proc MUC-6*, Columbia MD.

Sutton C, McCallum A (2005). Composition of CRFs for transfer learning, *Proc HLT/EMNLP*, Vancouver.