Sources of Performance in CRF Transfer Training: a Business Name-tagging Case Study

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

Jonathan Huggins
The MITRE Corporation
202 Burlington Rd.
Bedford, MA 01730 USA
jhuggins@mitre.org

Ben Wellner
The MITRE Corporation
202 Burlington Rd.
Bedford, MA 01730 USA
wellner@mitre,org

Abstract

This paper explores methods for increasing performance of CRF models, with a particular concern for transfer learning. We consider in particular the transfer case from political news to hard-to-tag business news, and show the effectiveness of several methods, including a novel semi-supervised approach.

Keywords

Entity extraction, machine learning, business intelligence.

1. Introduction: name tagging

Named entity recognition is one of the most widely studied problems in computational language processing. It was one of the first tasks to be treated with the corpus-based method, and has remained a touchstone for benchmarking corpus-based algorithms and learning regimens. Part of the enduring interest is that name tagging continues to provide technical challenges that help drive research. In particular, while the fundamentals of training a name tagger are well understood, such barriers to practical application as robust coverage and transfer training remain active research areas.

Indeed, name-tagging systems tend to perform best when both training and test data are drawn from the same distribution of sources and sample times. However, even seemingly small divergences between training and test can lead to steep drop-offs in performance. Overcoming this lack of carry-over from training to test is thus a key precondition for practical entity recognition applications.

This paper addresses this issue in the context of training conditional random fields (CRFs) to tag named entities in business texts. We explore several orthogonal strategies for bringing a name tagger to bear on a new domain, with the aim of providing high test-time performance, robustness to out-of-training phenomena, and minimal transfer training costs. We apply these strategies to a business news corpus, and achieve substantial performance gains while only requiring modest investments in transfer training.

2. Tagging business news

The potential divergence between a name tagger's training and test performance was documented as far back as the MUC7 evaluation [15]. At issue was a shift in topic between training and test conditions: from air incidents to satellite launches. While the training and test data were otherwise comparable (same sources, same broad topic of aerospace), several system developers implicated this as a cause for poor test-time performance.

In a recent study [18], we sought to quantify this divergence in the case of business texts. Our study found a substantial training-to-test performance gap for several mature recent systems trained (or hand-configured) to process current events news. While many of the systems did well with current events, their F scores dropped by 15 to 25 points for business news and financial reports.

The present paper takes these observations as a challenge to train a business entity tagger. The business genre is primarily of interest here as a case study, though all the more interesting because it appears so challenging. The framework we have chosen towards this end is that of conditional random fields. In the few years since their introduction [10], CRF models have enjoyed a groundswell of interest, especially as a method for discriminative sequence labeling. They have been applied to conventional sequence labeling tasks like part-of-speech tagging [20] or chunking [14], and unconventional ones like anonymization [21].

For our purposes, conditional random fields provide a number of distinct advantages. A key factor is that discriminative CRF training is not confounded by conditionally dependent features. This makes it safe to include useful features that may be conditionally dependent, *e.g.* lexical and part-of-speech n-grams. This also allows for features that encode non-local dependencies and external knowledge sources: these typically capture generalizations that co-vary in useful ways with the data, and are thus not independent of other features. CRF training also scales well, even with large numbers of n-gram features.

Finally, a CRF allows for post-hoc adjustment of the prior probability of a label. By artificially decreasing the prior, one causes the CRF decoder to generate fewer instances of the label, hence increasing precision at the expense of recall [12]; this proved very useful in this work.

We used the Carafe open source CRF package.[1]

[1] Available at http://sourceforge.net/projects/carafe

International Conference RANLP 2009 - Borovets, Bulgaria, pages 465–470

Source	Token count	Description
Reuters M+A	31,000 train 33,000 dev test 47.500 eval	Mergers and acquisition
Reuters BN	22,500 train 26.500 eval	General business news
Reuters HS	15,000 eval	Hot stocks
Reuters NI	6,400 eval	New Initiatives
NYT	78,500 train	General biz.
MUC 6	153,000 train*	Political news

Table 1: data samples (* = previously annotated)

3. Outline of our experiments

The experimental conditions we investigated fall roughly into two orthogonal types of considerations: training data (on the one hand) and features (on the other). For expository purposes, we have organized these experiments into four groupings, each representing a source of performance in training a CRF-based business entity extraction system.

- Cross-training: transfer learning experiments that exploit existing MUC6 data.
- Nearly unsupervised training: supervised training based on machine-generated data.
- Non-local knowledge: various strategies based on gazetteer and found name lists.
- Validation: evaluation with other kinds of business data

We found that each of these training conditions provided an increase in performance. Most interesting, these increases were largely independent, so the performance gain achieved by combining all these methods was essentially the sum of their independent performance gains.

3.1 Experimental data

The majority of our data was drawn from the Reuters business page. These data are plentiful and easy to harvest: we used an ad-hoc Web crawler to spider several business topics, and collected news stories for select time periods in 2006 and 2007. We manually annotated a small portion of these data according to the MUC6 standard (which calls for person, organization, location, date, time, money, and percent entities). We also collected and annotated a small comparable corpus of New York Times business stories. We did not reuse the data from our earlier study primarily because the samples were small and not wholly consistent.

Table 1 shows the annotated data samples we ended up using. Our training sample consisted of all the stories from the Reuters merger and acquisition topic (M+A) on March 5, 2007. Day-to-day (dev-test) scoring was conducted with the M+A stories from February 28. A final post-development round of evaluation was run with the chronologically distant M+A stories from June 21.

The other Reuters samples (BN, HS, and NI) cover business topics distinct from M+A. We used the February 28 stories from these topics to assess the generalization of the M+A models to related but off-topic news. For our baseline-setting and first cross-training runs, we used the original MUC6 training set; the second BN and NYT samples were used as additional cross-training data. Finally, we used two entire months of M+A data (November 2006 and May 2007) for our nearly unsupervised training runs.

Train	M+A dev test		M+A eval test	
	F	Δ error	F	Δ error
MUC	75.75	—	77.44	—
MUC + M+A	89.13	-55.2%	90.86	-59.5%
M+A	87.98	—	90.40	—
M+A + MUC	89.13	-9.6%	90.86	-4.8%

Table 2: baseline and cross-training scores.

3.2 Experimental set-up

Prior to either training or testing, texts were sentence-tagged and tokenized, and then given part-of-speech tags by a revised version of the Brill tagger [1]. We included conventional-case headlines in the Reuters stories, but excluded headline-case headers from NYT and MUC6. Training was through log-likelihood learning, with LBFG-S optimization and regularization with a Gaussian prior.

We used the MUC scorer for evaluation, as it allows for comparisons to the original MUC evaluations and to our earlier study. Note that the MUC scorer gives partial credit when system responses match the answer key in extent but not type (or vice versa), yielding somewhat higher scores than the popular CoNLL scorer.

4. Cross-training

In practical applications of entity extraction, it is commonly the case that standard training sets do not align exactly with the data of interest. For business news, our earlier study showed that the widely available MUC data set does not by itself provide adequate training to capture the entity distributions and writing style of the business pages [18]. One common piece of folk wisdom for this situation suggests pairing a modest sample of task-specific data with one of the common large data sets (MUC, ACE or CoNLL). If the two data sets are reasonably consistent, we would expect the larger corpus to contain relevant training instances that provide value beyond training on the task sample alone.

In this first set of experiments, we considered this simple form of transfer learning by pairing our M+A sample with the MUC6 corpus. Table 2 summarizes our results. Training on MUC6 alone produced an uninspiring M+A dev test score of F=75.75, which is consistent with our earlier study. The combination training yielded a score of F=89.13, for a 55% reduction in the error term. The combination also outperformed training on the genre-specific M+A data alone (F=87.98), which confirms the expectation that training transfer is taking place across the genres. As the table shows, this pattern also held true for our eval test.

This approach corresponds to Daumé's all-data transfer learning case [4]. Because the transfer in our case takes place between data annotated to the exact same standard, we tend to think of it as cross-training. Though effective, this approach is less general than transfer learning efforts that seek to leverage existing data sets against data with divergent entity types, i.e., different repertoires of entities or inconsistent definitions of their common entities [16].

5. Nearly unsupervised training

Our second set of experiments aims at increasing the volume of task-specific training data (always a good thing) without requiring substantial manual annotation. Indeed, since manual annotation can be costly and time-consuming, it is important to maximize the effectiveness of annotation efforts. Tried-and-true approaches towards this end have sought to increase annotator productivity through mixed-initiative bootstrapping. Typically, a model gets built from a small initial annotated corpus, with the model then guiding subsequent annotation, either through pre-tagging [5] or by screening training cases, as in active learning [3].

Although these techniques can substantially speed up annotation, they still require an annotator to read and validate every training instance. The alternative we consider here focuses on finding large numbers of training instances with only minimal manual intervention. The basic strategy is to locate these training instances in untagged text by a high-precision (nearly) automatic method. Given a large supply of untagged texts, the instance finder need only have modest recall in order to produce a useful training corpus.

5.1 Identifying company names in Reuters

We were able to devise this kind of instance-finding scheme for company names in Reuters business news. We rely on the fact that in some 5-10% of cases, company names are marked with a stock market ticker symbol, as in:

- *Bear Stearns Cos. (BSC.N ...)*

Our instance finder identifies these ticker forms through regular expressions, and then labels the sequence of capitalized words to the left of the form as a company name. Since companies represent the most frequent entity type in business news, and are also the hardest to tag, we hoped this scheme would automatically provide us many more training instances for precisely the most critical cases.

Some subtleties preclude this method from being entirely automatic. In particular, company names regularly include prepositions, conjunctions, or punctuation, as in:

- *"Helen of Troy Ltd."*
- *"JP Morgan Chase and Co."*
- *"Wong's Kong King (Holdings) Ltd."*

The instance finder must capture these non-noun atoms in such cases, but must also exclude them in others, as in:

- *"Jeff Schuman of Keefer Bruyette Woods."*

The instance finder must also exclude non-name modifiers that happen to be capitalized at the start of a sentence:

- *"Bootmaker Timberland Cos."*

To prevent the instance finder from producing incomplete names (e.g. "Troy Ltd.") or overly long ones ("Bootmaker …"), we included an as-needed manual review of potentially problematic contexts. The instance finder detects these contexts automatically, and after review, valid instances are cached so that they need not be queried again. Likewise predictive pre-nominals identified in the review ("bootmaker") are thereafter automatically removed from sentence-initial cases. Finally, to increase yield, the instance finder re-analyzes each story, looking for further mentions of found names. Mentions duplicating the names in their entirety or in shortened form are also labeled as companies. Except for a few easily-identified cases (e.g., "Ford" … "Ford of Canada"), this requires no review.

This mixed-initiative strategy proved highly effective, achieving precision of P=99.9 on large samples of Reuters news, for a recall of R=38 (measured on our M+A data). In addition, caching greatly reduced the need for annotator intervention. Once the cache got going, to process an entire month of M+A news (over 1,700 stories) required 50-100 interventions over 20 minutes or less. In comparison, full manual annotation of a single weekday of M+A data required several days of effort from experienced annotators.

5.2 Partial annotation and complete sentences

A salient property of this nearly unsupervised markup is that it is partial. So while the entities reported by the instance finder are essentially always accurate, the 38% recall level still leaves another 62% of organizations unreported – to say nothing of other entity types such as persons, places, money, and the like. These unreported entities make it hard to use instance finder data directly for supervised training. The issue is that the exact same entity may appear both as a positive example (that the instance finder found) and as a negative one (that it missed). This effectively causes the training procedure to ignore both examples, thus diminishing the potential contribution of the instance finder output. While a number of researchers have made strides recently towards semi-supervised learning, where not all data are annotated (*e.g.* [11]), these approaches typically pair a fully supervised corpus with separate annotated data, and do not speak to the case of partially annotated data.

The approach we used here was to sub-select those instance finder sentences that we could guarantee to be in fact completely annotated. Ignoring numeric entities for now (dates, money, …), a reliable gauge of complete annotation is the absence of any un-accounted-for capitalized words. We consider a capitalized word to be accounted for if either (i) it is labeled by the instance finder, or (ii) it is a closed-class word in sentence-initial position. In our trials, 8% to 9% of sentences in the corpus meet this criterion, yielding a substantial sub-corpus of fully-annotated sentences.

Train	M+A dev test		M+A eval test	
	F	**Δ error**	**F**	**Δ error**
M+A	87.98	—	90.40	—
+ 1 month	88.59	-5.1%	91.51	-11.6%
+ 2 months	88.28	-2.5%	91.11	-7.4%
M+A + MUC	89.13	—	90.86	—
+ 1 month	90.48	-12.5%	92.52	-18.2%
+ 2 months	90.60	-13.5%	92.49	-17.8%

Table 3: performance of nearly-unsupervised training.

The problem is that this sub-corpus is highly skewed: it necessarily contains only instances of company names, as these are the only entities identified by the instance finder. To boost the representation of other entities, we trained a CRF to identify numeric forms, and another to identify persons and places. We then artificially lowered the Viterbi decoder priors for these two models, trading off recall for precision until the decoder effectively had 100% precision. As with the entity finder, recall is only modest (30% for persons, 58% for locations). Nevertheless, these CRFs accounted for many more capitalized strings, thereby yielding many more completely annotated sentences, now up to 17% to 19% of the corpus. By extrapolation from manually annotated data, we estimate this represents around a third of the sentences that actually contain entities.

5.3 Experimental results

Table 3 reports this approach's results using our M+A training data, the MUC6 development corpus, and several months' worth of complete sentences identified by the instance finder. For the dev test, adding one month of complete sentences to the base M+A corpus yielded a modest error reduction of 5.1%, while adding a second month resulted in performance drop, with the error reduction falling to 2.5%. Results on the eval test showed a similar pattern.

We were intrigued, however, that experiments that also used the MUC6 corpus yielded much better results. The addition of one or two months of complete sentences respectively yielded dev test error reductions of 12.5%, and 13.5%, with eval test reductions topping off at 18.2%. Why such better performance with MUC6? An analysis of dev test errors for the M+A plus two months case, showed that most of the new errors were spurious organizations, all of them capitalized words. The entity distributions for these data explain what happened (Figure 4). For the complete sentences, organization names are not just the most common entity type, but form a 52% majority. While the M+A corpus starts out with only 46% of its entities as organizations, adding more complete sentences eventually causes the proportion to top 50%. The increasing skew towards organizations eventually leads the model to assign an overly strong default organization label to any potential entity, i.e., to any capitalized word.

When starting from the M+A and MUC6 data, however, this effect is moderated by the fact that the proportion

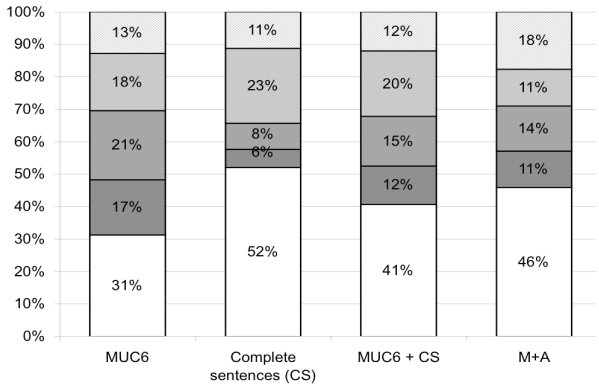

Figure 4: entity type distributions (from bottom to top: organization, person, location, money/percent, date/time)

of organizations in MUC6 is actually lower than it is in M+A. In effect the combination of MUC6 and the complete sentences yields an entity distribution that better matches that of the M+A data, hence providing greater opportunity to learn a high-performing model.

6. Non-local knowledge sources

As noted above, our post-hoc error analysis revealed cases where non-organizations were being labeled as organizations, among these person names and locations. Our third set of experiments attempted to address this problem by introducing non-local knowledge sources.

6.1 Gazetteer lists

Gazetteers of place and person names have long been used to improve the performance of name taggers [7]. We used lexical features to introduce gazetteers of given names, major geography, and numeric entity atoms (days, months, currencies). We avoided municipality lists, as they tend to also capture person names: of the 2,000 most common surnames in the US, most are also names of cities and towns.

6.2 Long-distance dependencies

One error that arose regularly in our post-hoc analysis concerned person names. The CRF generally identified full names like *"Thomas White"* but tended to mislabel surnames appearing on their own, *e.g. "White."* As Reuters avoids honorifics ("Mr.") the latter cases are hard to identify from context alone; a mechanism is thus needed to capture the implication between full names and bare surnames.

Various statistical approaches have been proposed to capture these long-distance name dependencies, *e.g.*, [2], [6], and especially [9], which presents a strategy based on the majority label for a word form. This approach proved effective for the CoNLL named entity task, so we reimplemented it here. As we detail elsewhere [19], the approach failed with our business data. Again, the prevalence of company names causes the CRF to assign the organization label to capitalized words when no countermanding

Features	M+A dev test		M+A eval test	
	F	**Δ error**	**F**	**Δ error**
M+A	87.98	—	90.40	—
+ gaz	89.27	-10.7%	92.07	-17.3%
+ LDD	88.74	-6.3%	91.58	-12.3%
+ gaz + LDD	90.23	-18.7%	92.82	-25.2%

Table 5: Effectiveness of non-local knowledge (NLK): gazetteers and long distance dependencies (LDD)

Train	M+A dev test		M+A eval test	
	F	**Δ error**	**F**	**Δ error**
M+A	87.98	—	90.40	—
M+A w/LDD	90.23	-18.7%	92.82	-25.2 %
+ 1 month	91.85	-32.2%	93.49	-32.2%
+ 2 months	91.32	-27.8%	93.06	-27.7%
M+A + MUC w/NLK	91.93	-32.9%	92.17	-18.4%
+ 1 month	93.21	-43.5%	93.95	-37.0%
+ 2 months	93.10	-42.6%	94.08	-38.3%

Table 6: Cross-training and nearly unsupervised training with non-local knowledge (NLK); Δ error is relative to M+A baseline.

evidence is available. In the case of person names, that evidence is primarily the presence of a given name, leading bare surnames to be mislabeled as organizations. Because bare surnames are more common than full names, the majority count strategy in [9] tends to perpetuate the error.

We adopted an alternative approach that captures long-distance name dependencies through evidence copying [19]. In the case of person names, if a word form α occurs in the right context of a given name, we copy this evidence to other instances of α through a non-local feature.

6.3 Experimental results

Using only our M+A development sample, we trained and evaluated CRF models with one or both of these knowledge sources active. Table 5 shows our results: while both knowledge sources were effective independently, it is particularly interesting that their combined application proved synergistic, yielding a greater performance gain than the sum of their separate yields.

Table 6 is even more telling. In this case, we activated the knowledge sources and repeated the cross-training and partially supervised training runs. The performance gains from these non-local knowledge sources were almost entirely independent of those produced by cross-training or the addition of complete sentences. Except for the final row in the table, which represents the largest training set, the non-local features contributed an additive performance boost. The configuration that performed highest on the dev test produced a compelling F score of 93.21 and an even higher F=93.95 on the eval test. Relative to training on M+A alone, this represents error reductions of 43.5% and 37% respectively.

7. Further validation runs

All of our experiments to this point were based on a single Reuters topic, mergers and acquisitions. To assess how well these M+A-trained models applied to business news in general, we annotated samples of several other business topics (previously shown in Table 1). The first two rows of Table 7 summarize performance of the M+A models on these test suites: for the most part, scores remained close to those measured on M+A data. Only the "hot stocks" topic showed degradation of more than around 1 point of F score.

Finally, to round off these cross-topic trials, we performed one more round of cross-training, adding in a sample of the BN topic along with editorially-dissimilar stories from the New York Times. The last row in Table 6 shows that cross-training was again effective at raising scores. As should be expected, performance on the BN eval test leapt higher, with a full 19% reduction in error. Interestingly, performance gains on the M+A eval test were not far behind, with a 16% reduction in error (6% for the dev test).

The most surprising gain however was with the lowest-performing HS topic, which gained over 2 points of F score, a 23% error reduction. Post-hoc analysis revealed why. The HS stories contain many references to the Dow Jones stock index (an index is not a company). In contrast. the M+A training data only had Dow Jones appearing as a company, thus leading the M+A model to mislabel HS references to the Dow as organizations. As the Times data happened to refer to the Dow as a non-name stock index, the cross-trained model removed the HS precision errors.

8. Discussion

We should begin by noting that our final F scores on blind eval data are comparable to those achieved by top-performing hand-built systems at the MUC6 and MUC7 evaluations. Further, this was achieved for business news, a genre that seems measurably harder to tag than the primarily political writing used in the MUC evaluations.

Training configuration	M+A dev	M+A eval	BN eval	HS eval	NI eval
M+A + MUC6 +1 month w/NLK	93.21	93.95	93.38	91.04	92.95
M+A + MUC6 +2 months w/NLK	93.10	94.08	93.12	90.56	93.06
M+A + MUC6 + BN + NYT +2 months w/NLK	93.48	95.05	94.41	92.72	93.47

Table 7: Performance on dissimilar topics

What is most interesting, however, is that this required so minimal an investment in new data annotation. While our highest scores were obtained using multiple new training sets, very respectable scores of F=93/94 (dev/test) were reached with only 31,000 words of newly annotated data, a roughly two-day annotation effort. We also did not need company lists: while these can improve recall, they are hard to keep current, and fail to capture most small businesses.

Key to our results is the roughly 40% error reduction provided by cross-training, nearly unsupervised training, and non-local knowledge. Among our core findings is that this combination of essentially orthogonal means yields an effectively additive reduction in error. This should be very encouraging to those seeking to apply entity extraction to new genres and new tasks.

It is interesting that highest performance required creative attention to both training regimens and knowledge sources. There is a methodological lesson here, as it is often tempting to focus research activities on only one or the other of these two threads.

One concern we would like to remediate in further work is the lack of direct comparison to recent efforts based on the CoNLL named entity scheme, including work on Hungarian business news [17]. The issue is complex: the CoNLL and MUC models differ not just as to scoring, but also around key annotation question.

Looking to the future, we are especially intrigued by the promise of our nearly unsupervised training strategy. While we used an instance finder that relied on the particulars of Reuters news, all that is required to apply the strategy is a high-precision instance finder with moderate recall. For the case of person names, for instance, honorifics like "Mr." act essentially like the ticker symbols we used in our company instance finder (we did not try this here because Reuters does not use honorifics). Another possibility is to generate instances through the application of gazetteers or known entity databases.

Unsupervised data have been used before as an adjunct to training for otherwise supervised entity extraction models. Approaches have included mutual bootstrapping [13] or self-training [8]. While these methods technically require no manual supervision, they tend to fail in unappealing ways once erroneous entities enter the self-generated training set. For this reason, the approach we've taken here, though requiring some manual review, may provide a valuable alternative in practice.

9. References

[1] Eric Brill. 1994. Some advances in rule-based part-of-speech tagging. Pcdgs. AAAI-94.

[2] Razvan Bunescu and Raymond J. Mooney. 2004. Collective information extraction with relational Markov networks. Pcdgs. of the 42nd ACL. Barcelona.

[3] Dagan, Ido and Sean Engelson. 1995. Committee-based sampling for training probabilistic classifiers. *Proc. ICML*.

[4] Daumé III, Hal. 2007. Frustratingly easy domain adaptation. *Proc. ACL*, Prague.

[5] Day, David, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patty Robinson, and Marc Vilain. 1997. Mixed-initiative development of language processing systems. *Proc. Applied ACL*.

[6] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sam-pling. Pcdgs. of the 43rd ACL. Ann Arbor, MI.

[7] Florian, Radu, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. *Proc. CoNLL.*

[8] Ji, Heng and Ralph Grishman. 2006. Data selection in semi-supervised learning for name tagging. *Proc. ACL*.

[9] Krishnan, Vijay, and Christopher Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. *Proc. ACL*, Sydney, Australia.

[10] Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proc. ICML*.

[11] Mann, Gideon and Andrew McCallum. 2007. Simple, robust, scalable semi-supervised learning via expectation regularization. *Proc. ICML*, Corvalis OR.

[12] Minkov E, Wang RC, Cohen WW. 2006. NER systems that suit user's preferences: adjusting the recall-precision trade-off for entity Extraction. *Proc HLT-NAACL*.

[13] Riloff, Elen and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. *Proc. AAAI.*

[14] Sha, Fei and Fernando Pereira. 2003. Shallow parsing with conditional random fields. *Proc. ACL-HLT.*

[15] Sundheim, Beth, *ed.* 1997. *Proc. of MUC-7*. Hosted at nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html

[16] Sutton, Chalres, and Andrew McCallum. 2005. Composition of conditional random fields for transfer learning. *Proc. HLT-EMNLP*, Vancouver, Canada.

[17] Szarvas, György, Richárd Farkas, László Felföldi, András Kocsor, & János Csirik. 2006. A highly accurate named entity corpus for Hungarian. *Proc. LREC*, Genoa.

[18] Vilain, Marc, Jennifer Su and Suzi Lubar. 2007. Entity extraction is a boring solved problem – or is it? *Proc NAACL.*

[19] Vilain, Marc, Jonathan Huggins, and Ben Wellner, 2009. Asimple feature-copying approach for long-distance dependecies. *Proc. CoNLL*, Boulder.

[20] Wellner, Ben and Marc Vilain. 2006. Leveraging machine-readable dictionaries in discriminative sequence models. *Proc. LREC*, Genoa,Italy.

[21] Wellner, Ben, Matt Huyck, Scott Mardis, John Aberdeen, Alex Morgan, Leon Peskin, Alex Yeh, Janet Hitzeman, Lynette Hirschman.2007. Rapidly retargetable approaches to de-identification. *J. Amer. Med. Informatics Assoc.* 14(5).