Redundancy-based Correction of Automatically Extracted Facts

Roman Yangarber and **Lauri Jokipii**
Department of Computer Science
University of Helsinki, Finland
first.last@cs.helsinki.fi

Abstract

The accuracy of event extraction is limited by a number of complicating factors, with errors compounded at all sages inside the Information Extraction pipeline. In this paper, we present methods for recovering automatically from errors committed in the pipeline processing. Recovery is achieved via post-processing facts aggregated over a large collection of documents, and suggesting corrections based on evidence external to the document. A further improvement is derived from propagating multiple, *locally non-best* slot fills through the pipeline. Evaluation shows that the global analysis is over 10 times more likely to suggest valid corrections to the local-only analysis than it is to suggest erroneous ones. This yields a substantial overall gain, with no supervised training.

1 Introduction

Information Extraction (IE) is a technology for finding facts in plain text, and coding them in a logical representation, such as, e.g., a relational database. IE is typically viewed and implemented as a sequence of stages—a "pipeline":

1. Layout, tokenization, lexical analysis
2. Name recognition and classification
3. Shallow (commonly,) syntactic parsing
4. Resolution of co-reference among entities
5. Pattern-based event matching and role mapping
6. Normalization and output generation

While accuracy at the lowest levels can reach high 90's, as the stages advance, complexity increases and performance degrades considerably.

The problem of IE as a whole, as well each of the listed subproblems, has been studied intensively for well over a decade, in many flavors and varieties. Key observations about much state-of-the-art IE are:

a. IE is typically performed by a pipeline process;

b. Only one hypothesis is propagated through the pipeline for each fact—the "best guess" the system can make for each slot fill;

c. IE is performed in a document-by-document fashion, applying *a priori* knowledge locally to each document.

The *a priori* knowledge may be encoded in a set of rules, an automatically trained model, or a hybrid thereof. Information extracted from documents—which may be termed *a posteriori* knowledge—is usually not reused across document boundaries, because the extracted facts are imprecise, and are therefore not a reliable basis for future extraction.

Furthermore, locally non-best slot fills are not propagated through the pipeline, and are consequently not available downstream, nor for any global analysis.

In most systems, these stages are performed in sequence. The locally-best slot fills are passed from the “lower-” to the “higher-level” modules, without feedback. Improvements are usually sought (e.g., the ACE research programme, (ACE, 2004)) by boosting performance at the lower levels, to reap benefits in the subsequent stages, where fewer errors are propagated.

The point of departure for this paper is: the IE process is noisy and imprecise at the single-document level; this has been the case for some time, and though there is much active research in the area, the situation is not likely to change radically in the immediate future—rather, we can expect slow, incremental improvements over some years.

In our experiments, we approach the performance problem from the opposite end: start with the extracted results and see if the totality of *a posteriori* knowledge about the domain—knowledge generated by the same noisy process we are trying to improve—can help recover from errors that stem from locally insufficient *a priori* knowledge.

The aim of the research presented in this paper is to improve performance by aggregating related facts, which were extracted from a large document collection, and to examine to what extent the correctly extracted facts can help correct those that were extracted erroneously.

The rest of the paper is organized as follows. Section 2 contains a brief review of relevant prior work. Section 3 presents the experimental setup: the text corpus, the IE process, the extracted facts, and what aspects of the the extracted facts we try to improve in this paper. Section 4 presents the methods for improving the quality of the data using global analysis, starting with a naive, baseline method, and proceeding with several extensions. Each method is then evaluated, and the results are examined in section 5. In section 6, we present further extensions currently under research, followed by the conclusion.

2 Prior Work

As we stated in the introduction, typical IE systems consist of modules arranged in a cascade, or a pipeline. The modules themselves are be based on heuristic rules or automatically trained, there is an abundance of approaches in both camps (and everywhere in between,) to each of the pipeline stages listed in the introduction.

It is our view that to improve performance we ought to depart from the traditional linear, pipeline-style design. This view is shared by others in the research community; the potential benefits have previously been recognized in several contexts.

In (Nahm and Mooney, 2000a; Nahm and Mooney, 2000b), it was shown that learning rules from a fact base, extracted from a corpus of job postings for computer programmers, improves future extraction, even though the originally extracted facts themselves are far from error-free. The idea is to mine the data base for association rules, and then to integrate these rules into the extraction process.

The baseline system is obtained by supervised learning from a few hundred manually annotated examples. Then the IE system is applied to successively larger sets of unlabeled examples, and association rules are mined from the extracted facts. The resulting combined system (trained model plus association rules) showed an improvement in performance on a test set, which correlated with the size of the unlabeled corpus.

In work on improving (Chinese) named entity tagging, (Ji and Grishman, 2004; Ji and Grishman, 2005), show benefits to this component from integrating decisions made in later stages, viz. coreference, and relation extraction.[1]

Tighter coupling and integration between IE and KDD components for mutual benefit is advocated by (McCallum and Jensen, 2003), which present models based on CRFs and supervised training.

This work is related in spirit to the work presented in this paper, in its focus on leveraging cross-document information that information—though it is inherently noisy—to improve local decisions. We expect that the approach could be quite powerful when these ideas are used in combination, and our experiments seem to confirm this expectation.

3 Experimental Setup

In this section we describe the text corpus, the underlying IE process, the form of the extracted facts, and the specific problem under study—i.e., which aspects of these facts we first try to improve.

[1] Performance on English named entity tasks reaches mid to high 90’s in many domains.

3.1 Corpus

We conducted experiments with redundancy-based auto-correction over a large database of facts extracted from the texts in ProMED-Mail, a mailing list which carries reports about outbreaks of infectious epidemics around the world and the efforts to contain them. This domain has been explored earlier; see, e.g., (Grishman et al., 2003) for an overview.

Our underlying IE system is described in (Yangarber et al., 2005). The system is a hybrid automatically- and manually-built pattern base for finding facts, an HMM-based name tagger, automatically compiled and manually verified domain-specific ontology, based in part on MeSH, (MeS, 2004), and a rule-based co-reference module, that uses the ontology.

The database is live on-line, and is continuously updated with new incoming reports; it can be accessed at `doremi.cs.helsinki.fi/plus/`.

Text reports have been collected by ProMED-Mail for over 10 years. The quality of reporting (and editing) has been rising over time, which is easy to observe in the text data. The distribution of the data, aggregated by month is shown in Figure 1, where one can see a steady increase in volume over time.[2]

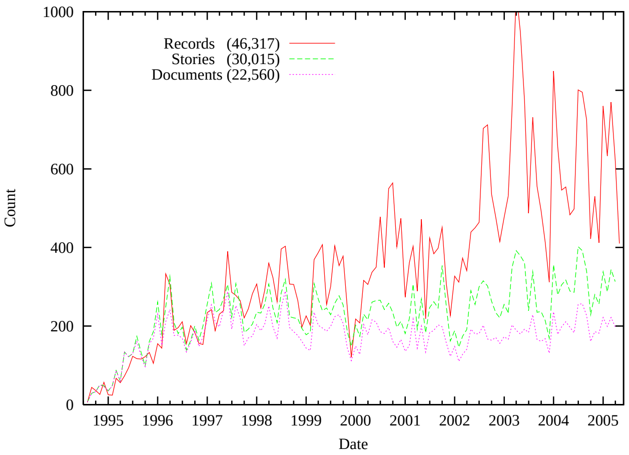

Figure 1: Distribution of data in ProMED-Mail

3.2 Extracted Facts

We now describe the makeup of the data extracted from text by the IE process, with basic terminology.

Each document in the corpus, contains a single *report*, which may contain one or more *stories*. Story breaks are indicated by layout features, and are extracted by heuristic rules, tuned for this domain and corpus. When processing a multi-story report, the IE system treats each story as a separate document; no information is shared among stories, except that the text of the main headline of a multi-story report is available to each story. [3]

Since outbreaks may be described in complex ways, it is not obvious how to represent a single fact in this context. To break down this problem, we use the notion of an *incident*. Each story may contain multiple outbreak-related incidents/facts.[4]

We analyze an outbreak as a series of incidents. The incidents may give "redundant" information about an outbreak, e.g., by covering overlapping time intervals or geographic areas. For example, a report may first state the number of cases within the last month, and then give the total for the entire year. We treat each of these statements as a separate incident; the containment relations among them are beyond the scope of our current goals.[5]

Thus each incident corresponds to a partial description of an outbreak, over a period of time and geographic area. This makes it easy to represent each incident/fact as a separate row in the table.

The key fields of the incident table are:

- Disease Name
- Location
- Date (start and end)

Where possible, the system also extracts information about the victims affected in the incident—their count, severity (affected or dead), and a descriptor (people, animals, etc.). The system also extracts bookkeeping information about each incident: locations of mentions of the key fields in the text, etc.

The system's performance is currently at 71.16 F-measure: 67% recall, 74% precision. This score is obtained by a MUC scorer (Douthat, 1998) on a 50-document test corpus, which was manually tagged with correct incidents with these slots. We have

[2] This is beneficial to the IE process, which operates better with formulaic, well-edited text.

[3] The format of the documents in the archive can be examined by browsing the source site `www.promedmail.org`.

[4] In this paper, we use the terms *fact*, *incident*, and *event* interchangeably.

[5] This problem is addressed in, e.g., (Huttunen et al., 2002).

no blind-test corpus at present, but prior experience suggests that we ought to expect about a 10% reduction in F-measure on unseen data; this is approximately borne out by our informal evaluations.

Further, the system attempts to "normalize" the key fields. An alias for a disease name (e.g., "bird flu") is mapped to a *canonical* name ("avian influenza.")[6] Date expressions are normalized to a standard format `yyyy.mm.dd–yyyy.mm.dd`.[7]

Note that the system may not be able to normalize some entities, which then remain un-normalized.

Such normalization is clearly helpful for searching, but it is not only a user-interface issue. Normalizing reduces sparseness of data; and since our intent is to aggregate related facts across a large fact base, excessive variation in the database fields would reduce the effectiveness of the proposed methods.

3.3 Experimental Focus: Location Normalization

A more complex problem arises out of the need to normalize location names. For each record, we normalize the location field—which may be a name of a small village or a larger area—by relating it to the name of the containing country; we also decided to map locations in the United States to the name of the containing state, (rather than the name of the country, "USA").[8] This mapping will be henceforth referred to as "location–state," for short. The ideas presented in the introduction are explored in the remainder of this paper in the context of correcting the location–state mapping.

Section 6 will touch upon our current work on extending the methodology to slots other than *state*. (Please see Section 5 for further justification of this choice for our initial experiments.)

To make the experiments interesting and fair, we kept the size of the gazetteer small. The *a priori* geographic knowledge base contains names of countries of the world (270), with aliases for several of them; a list of capitals and other selected major cities (300); a list of states in the USA and acronyms (50); major US cities (100); names of the (sub)continents (10), and oceans. In our current implementation, continents are treated semantically as "states" as well.[9]

The IE system operates in a local, document-by-document fashion. Upon encountering a location name that is not in its dictionaries, the system has two ways to map it to the state name. One way is by matching patterns over the immediate local context, ("Milan, Italy"). Failing that, it tries to find the corresponding state by positing an "underspecified" state name (as if referred to by a kind of special "pronoun") and mapping the location name to that. The reference resolution module then finds the most likely antecedent entity, of the semantic type "state/country," where likelihood is determined by its proximity to the mention of the location name.

Note that the IE system outputs only a single, best hypothesis for the *state* fill for each record.

3.4 The Data

The database currently contains about $46,317$ individual facts/incidents, extracted from $30,015$ stories, from $22,560$ reports (cf. Fig. 1). Each incident has a *location* and a *state* filler. We say a location name is "ambiguous" if it appears in the *location* slot of at least two records, which have different names in the *state* slot. The number of distinct "ambiguous" location names is $1,020$.

Note, this terminology is a bit sloppy: the fillers to which we refer as "ambiguous location names", may not be valid location names at all; they may simply be errors in the IE process. E.g., at the name classification stage, a disease name (especially if not in the disease dictionary) may be misclassified, and used as a filler for the *location* slot.

We further group together the *location* fills by stripping lower-case words that are not part of the proper name, from the front and the end of the fill. E.g., we group together "southern Mumbai" and "the Mumbai area," as referring to the same name.

After grouping and trimming insignificant words, the number of distinct names appearing in *location* fills is 600, which covers a total of 6583 records, or 14.2% of all extracted facts. As an estimate of the *potential* for erroneous mapping from locations to states, this is quite high, about 1 in 7 records.[10]

[6] This is done by means of a set of scenario-specific patterns and a dictionary of about 2500 disease names with aliases.

[7] Some date intervals may not have a starting date, e.g., if the text states "As of last Tuesday, the victim count is N..."

[8] This decision was made because otherwise records with state = USA strongly skew the data, and complicate learning.

[9] By the same token, both *Connecticut* and *USA* are "states."

[10] Of course, it can be higher as well, if the IE system con-

4 Experiments and Results

We now present the methods of correcting possible errors in the location–state relation. A method M tries to suggest a new value for the *state* fill for every incident **I** that contains an ambiguous *location* fill:

$$NewState(\mathbf{I}) = \arg\max_{s \in \mathcal{S}_M(\mathbf{I})} Score_M(s, \mathbf{I}) \quad (1)$$

where $\mathcal{S}_M(\mathbf{I})$ is a set of all candidate states considered by M for **I**; $Score_M(s, \mathbf{I})$ is the scoring function specific to M. The method chooses the candidate state which maximizes the score.

For each method below, we discuss how $\mathcal{S}_M$ and $Score_M$ are constructed.

4.1 Baseline: Raw Majority

We begin with a simple recovery approach. We simply assume that the correct state for an ambiguous location name is the state most frequently associated with it in the database. We denote by $\mathcal{D}$ the set of all incidents in the database. For an incident $\mathbf{I} \in \mathcal{D}$, we write $l, s \sqsubset \mathbf{I}$ when location l, state s, etc., "belong" to **I**, i.e., are extracted as fills in **I**. In the baseline method, B, for each incident **I** where $l \sqsubset \mathbf{I}$ is one of the 600 ambiguous location names, we define:

$$\mathcal{S}_B(\mathbf{I}) = \{s' : \exists \mathbf{I}' \in \mathcal{D} \mid (l, s') \sqsubset \mathbf{I}'\}$$
$$Score_B(s', \mathbf{I}) = |\{\mathbf{I}' \in \mathcal{D} \mid (l, s') \sqsubset \mathbf{I}'\}|$$

i.e., s' is a candidate if it is a *state* fill in some incident whose *location* fill is also l; the score is the number of times the pair (l, s') appear together in some incident in $\mathcal{D}$. The majority winner is then suggested as the "correct" state, for *every* record containing l. By "majority" winner we mean the candidate with the *maximal* count, rather than a state with more than half of the votes. When the candidates tie for first place, no suggestions are made—although it is quite likely that some of the records carrying l will have incorrect *state* fills.

A manual evaluation of the performance of this method is shown in Table 1, the *Baseline* column.

The first row shows for how many records this method suggested a change from the original, IE-filled *state*. The baseline changed 858 incidents. This constitutes about 13% out of the maximum number of changeable records, 6583.

Thus, this number represents the volume of the potential gain or loss from the global analysis: the proportion of records that actually get modified.

The remaining records were unchanged, either because the majority winner coincides with the original IE-extracted state, or because there was a tie for the top score, so no decision could be made.

We manually verified a substantial sample of the modified records. When verifying the changes, we referred back to the text of the incident, and, when necessary, consulted further geographical sources to determine exactly whether the change was correct in each case.

For the baseline we had manually verified 27.7% of the changes. Of these, 68.5% were a clear gain: an incorrect state was changed to a correct state. 6.3% were a clear loss, a correct state lost to an incorrect one. This produces quite a high baseline, surprisingly difficult to beat.

The next two rows represent the "grey" areas. These are records which were difficult to judge, for one of two technical reasons. A: the "location" name was itself *erroneous*, in which case these redundancy-based approaches are not meaningful; or, B: the suggestion replaces an area by its subregion or super-region, e.g., changing "Connecticut" to "USA", or "Europe" to "France."[11]

Although it is not strictly meaningful to judge whether these changes constitute a gain or a loss, we nonetheless tried to assess whether changing the state hurt the accuracy of *the incident*, since the incident may have a correct state even though its location is erroneous (case A); likewise, it may be correct to say that a given location is indeed a part of Connecticut, in which case changing it to USA loses information, and is a kind of loss.

That is the interpretation of the grey gain and loss instances. The final row, *no loss*, indicates the proportion of cases where an originally incorrect state name was changed to a new one, also incorrect.

sistently *always* maps some location name to the same wrong state; these cases are below the radar of our scheme, in which the starting point is the "ambiguous" locations.

[11] Note, that for some locations, which are not within any one state's boundary, a continent is a "correct state", for example, "the Amazon Region," or "Serengeti."

Records	*Baseline*		*DB-filtered*		*Confidence*		*Multi-candidate*	
Changed	13.0%	$858/6583$	8.7%	$577/6583$	9.7%	$642/6583$	$\mathbf{16.2}\%$	$1072/6583$
Verified	27.7%	$238/858$	38.6%	$223/577$	37.1%	$238/642$	26.5%	$284/1072$
Gain	68.5%	$163/238$	71.3%	$159/223$	80.3%	$191/238$	$\mathbf{76.4}\%$	$217/284$
Loss	6.3%	$15/238$	2.2%	$5/223$	1.3%	$3/238$	$\mathbf{3.5}\%$	$10/284$
Grey gain	10.9%	$26/238$	12.6%	$28/223$	11.8%	$28/238$	13.7%	$39/284$
Grey loss	6.7%	$16/238$	5.4%	$12/223$	0.0%	$0/238$	2.1%	$6/284$
No loss	7.6%	$18/238$	8.5%	$19/223$	6.7%	$16/238$	4.2%	$12/284$

Table 1: Performance of Correction Methods

4.2 Database Filtering

Next we examined a variant of baseline raw majority vote, noting that simply choosing the state most frequently associated with a location name is a bit naive: the location–state relation is not functional—i.e., some location names map to more than one state in reality. There are many locations which share the same name.[12]

To approach this more intelligently, we define:

$$\mathcal{S}_F(\mathbf{I}) = \mathcal{S}_B(\mathbf{I}) \cap StatesInStory(\mathbf{I})$$
$$Score_F(s', \mathbf{I}) = Score_B(s', \mathbf{I})$$

The baseline vote counting across the data base (DB) produced a ranked list of candidate states s' for the location $l \sqsubset \mathbf{I}$. We then filtered this list through $StatesInStory(\mathbf{I})$, the list of states mentioned in the story containing the incident $\mathbf{I}$. The filtered majority winner was selected as the suggested change.

For example, the name "Athens" may refer to the city in Greece, or to the city in Georgia (USA). Suppose that Greece is the raw majority winner. The baseline method will always tag all instances of Athens as being in Greece. However, in a story about Georgia, Greece will likely not be mentioned at all, so it is safe to rule it out. This helps a minority winner, when the majority is not present in the story.

Surprisingly, this method did not yield a substantial improvement over the baseline, (though it was more careful by changing fewer records). This may indicate that NWP is not an important source of errors here: though many truly ambiguous locations do exist, they do not account for many instances in this DB.

4.3 Confidence-Based Ranking

A more clear improvement over the baseline is obtained by taking the *local confidence* of the state–location association into account. For each record, we extend the IE analysis to produce a confidence value for the state. Confidence is computed by simple, document-local heuristics, as follows:

If the location and state are both within the span of text covered by the incident—text which was actually matched by a rule in the IE system,—or if the state is the *unique* state mentioned in the story, it gets a score of 2—the incident has high confidence in the state. Otherwise, if the state is outside the incident's span, but is inside the same sentence as the incident, and is also the unique state mentioned in that sentence, it gets a score of 1. Otherwise it receives a score of zero.

Given the confidence score for each (location l, state s) pair, the majority counting is based on the *cumulative confidence*, $conf_{state}$ (l, s) in the DB, rather than on the *cumulative count* of occurrences of this pair in the DB:

$$\mathcal{S}_C(\mathbf{I}) = \mathcal{S}_F(\mathbf{I})$$
$$Score_C(s', \mathbf{I}) = \sum_{\mathbf{I'} \in \mathcal{D} | (l, s') \sqsubset \mathbf{I'}} conf_{state}(\mathbf{I'})$$

Filtering through the story is also applied, as in the previous method. The resulting method favors more correct decisions, and fewer erroneous ones.

We should note here, that the notion of confidence of a fill (here, the state fill) is naturally extended to the notion of confidence of a *record*: For each of

[12] We refer to this as the "New-World phenomenon" (NWP), due to its prevalence in the Americas: "Santa Cruz" occurs in several Latin American countries; locations named after saints are common. In the USA, city and county names often appear in multiple states—Winnebago County, Springfield; many cities are named after older European cities.

the three key fills—location, date, disease name—compute a confidence based on the same heuristics. Then we say that a record $\mathbf{I}$ has high confidence, if it has non-zero confidence in all three of the key fills. The notion of record confidence is used in Section 6.

4.4 Multi-Candidate Propagation

Finally, we tried propagating multiple candidate state hypotheses for each instance of an ambiguous location name l:

$$\mathcal{S}_+(\mathbf{I}) = \bigcup_{\mathbf{I}' \in \mathcal{D} | l \sqsubset \mathbf{I}'} StatesInStory(\mathbf{I}')$$

$$Score_+(s', \mathbf{I}) = \sum_{\mathbf{I}' \in \mathcal{D} | l \sqsubset \mathbf{I}'} prox(s', \mathbf{I}')$$

where the proximity is inversely proportional to the distance of s' from incident $\mathbf{I}'$, in the story of $\mathbf{I}'$:

$$prox(s, \mathbf{I}) = \begin{cases} \frac{1}{Z(\mathbf{I})} \cdot \frac{1}{\Delta(s, \mathbf{I}) + 1} & if\ s \sqsubset \mathbf{I} \\ 0 & otherwise \end{cases}$$

For an incident $\mathbf{I}$ mentioning location l, the IE system outputs the list of all states $\{s\}$ mentioned in the same story; we then rank each s according to the inverse of *distance* Δ: the number of sentences between $\mathbf{I}$ and s. $Z(\mathbf{I})$ is a normalization factor.

The proximity for each pair (l, s), is between 0 and 1. Rather than giving a full point to a single, locally-best guess among the s's, this point is shared proportionately among all competing s's. For example, if states s_0, s_1, s_5 are in the same sentence as $\mathbf{I}$, one, and five sentences away, respectively, then $Z(\mathbf{I}) = 1 + \frac{1}{2} + \frac{1}{6} = \frac{5}{3}$, and $prox(s_0) = 1 \cdot \frac{3}{5} = \frac{3}{5}$, $prox(s_1) = \frac{1}{2} \cdot \frac{3}{5} = \frac{3}{10}$,and $prox(s_5) = \frac{1}{6} \cdot \frac{3}{5} = \frac{1}{10}$.

The score for each state s for the given l is then the sum of proximities of s to l across all stories.

The resulting performance is substantially better than the baseline, while the *number of changed* records is substantially higher than in the competing methods. This is due to the fact that this method allows for a much larger pool of candidates than the others, and assigns to them much smoother weights, virtually eliminating ties in the ranking among hypotheses.

5 Discussion

Among the four competing approaches presented above, the baseline performs surprisingly well. We should note that this research is not aimed specifically at improving geographic Named Entity resolution. It is the first in a series of experiments aiming to leverage redundancy across a large fact base extracted from text, to improve the quality of extracted data. We chose to experiment with this relation first because of its simplicity, and because the *state* field is a key field in our application.

For this reason, the *a priori* geographic knowledge base was intentionally not as extensive as it might have been, had we tried in earnest to match locations with corresponding states (e.g., by incorporating the CIA Factbook, or other gazetteer).

The intent here is to investigate how a relation can be improved by leveraging redundancy across a large body of records. The support we used for geographic name resolution was therefore deliberately modest, cf. Section 3.3.

It is quite feasible to enumerate the countries and the larger regions, since they number in the low hundreds, whereas there are many tens of thousands of cities, towns, villages, regions, districts, etc.

6 Current Work

Three parallel lines of current research are:

1. combining evidence from multiple features
2. applying redundancy-based correction to other fields in the database
3. back-propagation of corrected results, to repair components that induced incorrect information.

The results so far presented show that even a naive, intuitive approach can help correct local errors via global analysis. We are currently working on more complex extensions of these methods.

Each method exploits one main feature of the underlying data: the distance from candidate state to the mention of the location name. In the multi-candidate hypothesis method, this distance is exploited explicitly in the scoring function. In the other methods, it is used inside the co-reference module of the IE pipeline, to find the (single) locally-best state.

However, other textual features of the state candidate should contribute to establishing the relations to a location mention, besides the raw distance. For example, at a given distance, it is very important whether the state is mentioned before the location (more likely to be related) vs. after the location (less likely). Another important feature: is the state mentioned in the main story/report headline? If so, its score should be raised. It is quite common for documents to declaim the focal state only once in the headline, and never mention it again, instead mentioning other states, neighboring, or otherwise relevant to the story. The distance measure used alone may be insufficient in such cases.

How are these features to be combined? One path is to use some combination of features, such as a weighted sum, with parameters trained on a manually tagged data set. As we already have a reasonably sized set tagged for evaluation, we can split it into two, train the parameter on a larger portion, evaluate on a smaller one, and cross-validate.

We will be using this approach as a baseline. However, we aim to use a much larger set of data to train the parameters, without manually tagging large training sets.

The idea is to treat the set of incidents with high *record confidence*, Sec. 4.3, rather than manually tagged data, as ground truth. Again, there "confident" truth will not be completely error-free, but because error rates are lower among the confident records, we may be able to leverage global analysis to produce the desired effect: training parameters for more complex models—involving multiple features—for global re-ranking of decisions.

Conclusion

Our approach rests on the idea that evidence aggregated across documents should help resolve difficult problems at the level of a given document.

Our experiments confirm that aggregating global information about related facts, and propagating locally non-best analyses through the pipeline, provide powerful sources of additional evidence, which are able to reverse incorrect decisions, based only on local and *a priori* information.

The proposed approach requires no supervision or training of any kind. It does, however require a substantial collection of evidence across a large body of extracted records; this approach needs a "critical mass" of data to be effective. Although large volume of facts is usually not reported in classic IE experiments, obtaining high volume should be natural in principle.

Acknowledgement

We'd like to thank Winston Lin of New York University, who worked on an early version of these experiments.

References

2004. Automatic content extraction.

A. Douthat. 1998. The Message Understanding Conference scoring software user's manual. In *Proc. 7th Message Understanding Conf. (MUC-7)*, Fairfax, VA.

R. Grishman, S. Huttunen, and R. Yangarber. 2003. Information extraction for enhanced access to disease outbreak reports. *J. of Biomed. Informatics*, **35**(4).

S. Huttunen, R. Yangarber, and R. Grishman. 2002. Complexity of event structure in information extraction. In *Proc. 19th Intl. Conf. Computational Linguistics (COLING 2002)*, Taipei.

H. Ji and R. Grishman. 2004. Applying coreference to improve name recognition. In *Proc. Reference Resolution Wkshop, (ACL-2004)*, Barcelona, Spain.

H. Ji and R. Grishman. 2005. Improving name tagging by reference resolution and relation detection. In *Proc. ACL-2005*, Amherst, Mass.

A. McCallum and D. Jensen. 2003. A note on the unification of information extraction and data mining using conditional-probability, relational models. In *IJCAI'03 Workshop on Learning Statistical Models from Relational Data*.

2004. Medical subject headings.

U. Y. Nahm and R. Mooney. 2000a. A mutually beneficial integration of data mining and information extraction. In *AAAI-2000*, Austin, TX.

U. Y. Nahm and R. Mooney. 2000b. Using information extraction to aid the discovery of prediction rules from text. In *KDD-2000 Text Mining Wkshop*, Boston, MA.

R. Yangarber, L. Jokipii, A. Rauramo, and S. Huttunen. 2005. Extracting information about outbreaks of infectious epidemics. In *Proc. HLT-EMNLP 2005 Demonstrations*, Vancouver, Canada.