Improving Mention Detection Robustness to Noisy Input

Radu Florian, John F. Pitrelli, Salim Roukos and Imed Zitouni
IBM T.J. Watson Research Center
Yorktown Heights, NY, U.S.A.
{raduf,pitrelli,roukos,izitouni}us.ibm.com

Abstract

Information-extraction (IE) research typically focuses on clean-text inputs. However, an IE engine serving real applications yields many false alarms due to less-well-formed input. For example, IE in a multilingual broadcast processing system has to deal with inaccurate automatic transcription and translation. The resulting presence of non-target-language text in this case, and non-language material interspersed in data from other applications, raise the research problem of making IE robust to such noisy input text. We address one such IE task: entity-mention detection. We describe augmenting a statistical mention-detection system in order to reduce false alarms from spurious passages. The diverse nature of input noise leads us to pursue a multi-faceted approach to robustness. For our English-language system, at various miss rates we eliminate 97% of false alarms on inputs from other Latin-alphabet languages. In another experiment, representing scenarios in which genre-specific training is infeasible, we process real financial-transactions text containing mixed languages and data-set codes. On these data, because we do not train on data like it, we achieve a smaller but significant improvement. These gains come with virtually no loss in accuracy on clean English text.

1 Introduction

Information-extraction (IE) research is typically performed on clean text in a predetermined language. Lately, IE has improved to the point of being usable for some real-world tasks whose accuracy requirements are reachable with current technology. These uses include media monitoring, topic alerts, summarization, population of databases for advanced search, etc. These uses often combine IE with technologies such as speech recognition, machine translation, topic clustering, and information retrieval.

The propagation of IE technology from isolated use to aggregates with such other technologies, from NLP experts to other types of computer scientists, and from researchers to users, feeds back to the IE research community the need for additional investigation which we loosely refer to as "information-extraction robustness" research. For example:

1. Broadcast monitoring demands that IE handle as input not only clean text, but also the transcripts output by speech recognizers.

2. Multilingual applications, and the imperfection of translation technology, require IE to contend with non-target-language text input (Pitrelli et al., 2008).

3. Naive users at times input to IE other material which deviates from clean text, such as a PDF file that "looks" like plain text.

4. Search applications require IE to deal with databases which not only possess clean text but at times exhibit other complications like markup codes particular to narrow, application-specific data-format standards, for example, the excerpt from a financial-transactions data set shown in Figure 1.

   Legacy industry-specific standards, such as illustrated in this example, are part of long-established processes which are cumbersome to convert to a more-modern database format. Transaction data sets typically build up over a period of years, and as seen here, can exhibit

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 335–345,
MIT, Massachusetts, USA, 9-11 October 2010.

```
:54D://121000358
BANK OF BOSTON
:55D:/0148280005
NEVADA DEPT.OF VET.94C RECOV.FD
-5:MAC:E19DECA8CHK:641EB09B8968

USING OF FIELD 59: ONLY /INS/ WHEN
FOLLOWED BY BCC CODE IN CASE
OF QUESTIONS DONT HESITATE TO
CONTACT US QUOTING REFERENCE
NON-STC CHARGES OR VIA E-MAIL:
YOVANKA(UL)BRATASOVA(AT)BOA.CZ.
BEST REGARDS
BANKA OBCHODNIKA, A.S. PRAGUE, CZ

:58E::ADTX//++ ADDITIONAL
INFORMATION ++ PLEASE BE
INFORMED THAT AS A RESULT OF
THE PURCHASE OFFER ENDED ON 23
MAR 2008 CALDRADE LTD. IS
POSSESSING WITH MORE THEN 90
PER CENT VOTING RIGHT OF SLICE.
THEREFOR CALDRADE LTD. IS
EXERCISING PURCHASE RIGHTS
FOR ALL SLICE SHARES WHICH ARE
CURRENTLY NOT INHIS OWN.
PURCHASE PRICE: HUF 1.940 PER
SHARE. PLEASE :58E::ADTX//NOTE
THAT THOSE SHARES WHICH WILL
NOT BE PRESENTED TO THE OFFER
WILL BE CANCELLED AND INVALID.

:58:SIE SELBST
TRN/REF:515220 035
:78:RUECKGABE DES BETRAGES LT.
ANZBA43 M ZWECKS RUECKGABE IN
AUD. URSPR. ZU UNSEREM ZA MIT
REF. 0170252313279065 UND IHRE
RUECKG. :42:/BNF/UNSERE REF:
```

Figure 1: Example application-specific text, in this case from financial transactions.

peculiar mark-up interspersed with meaningful text. They also suffer complications arising from limited-size entry fields and a diversity of data-entry personnel, leading to effects like haphazard abbreviation and improper spacing, as shown. These issues greatly complicate the IE problem, particularly considering that adapting IE to such formats is hampered by the existence of a multitude of such "standards" and by lack of sufficient annotated data in each one.

A typical state-of-the-art statistical IE engine will happily process such "noisy" inputs, and will typically provide garbage-in/garbage-out performance, embarrassingly reporting spurious "information" no human would ever mistake. Yet it is also inappropriate to discard such documents wholesale: even poor-quality inputs may have relevant information interspersed. This information can include accurate speech-recognition output, names which are recognizable even in wrong-language material, and clean target-language passages interleaved with the mark-up. Thus, here we address methods to make IE robust to such varied-quality inputs. Specifically, our overall goals are

- to skip processing non-language material such as standard or database-specific mark-up,
- to process all non-target-language text cautiously, catching interspersed target-language text as well as text which is compatible with the target language, *e.g.* person names which are the same in the target- and non-target language, and
- to degrade gracefully when processing anomalous target-language material,

while minimizing any disruption of the processing of clean, target-language text, and avoiding any necessity for explicit pre-classification of the genre of material being input to the system. Such explicit classification would be impractical in the presence of the interleaving and the unconstrained data formats from unpredetermined sources.

We begin our robustness work by addressing an important and basic IE task: mention detection (MD). MD is the task of identifying and classifying textual references to entities in open-domain texts. Mentions may be of type "named" (*e.g.* `John`, `Las Vegas`), "nominal" (*e.g.* `engineer`, `dentist`) or "pronominal" (*e.g.* `they`, `he`). A mention also

has a specific class which describes the type of entity it refers to. For instance, consider the following sentence:

```
Julia Gillard, prime
minister of Australia,
declared she will enhance
the country's economy.
```

Here we see three mentions of one person entity: `Julia Gillard`, `prime minister`, and `she`; these mentions are of type named, nominal, and pronominal, respectively. `Australia` and `country` are mentions of type named and nominal, respectively, of a single geopolitical entity. Thus, the MD task is a more general and complex task than named-entity recognition, which aims at identifying and classifying only named mentions.

Our approach to IE has been to use language-independent algorithms, in order to facilitate reuse across languages, but we train them with language-specific data, for the sake of accuracy. Therefore, input is expected to be predominantly in a target language. However, real-world data genres inevitably include some mixed-language/non-linguistic input. Genre-specific training is typically infeasible due to such application-specific data sets being unannotated, motivating this line of research. Therefore, the goal of this study is to investigate schemes to make a language-specific MD engine robust to the types of interspersed non-target material described above. In these initial experiments, we work with English as the target language, though we aim to make our approach to robustness as target-language-independent as possible.

While our ultimate goal is a language-independent approach to robustness, in these initial experiments, English is the target language. However, we process mixed-language material including real-world data with its own peculiar mark-up, text conventions including abbreviations, and mix of languages, with the goal of English MD.

We approach robust MD using a multi-stage strategy. First, non-target-character-set passages (here, non-Latin-alphabet) are identified and marked for non-processing. Then, following word-tokenization, we apply a language classifier to a sliding variable-length set of windows in order to generate features for each word indicative of how much the text around that word resembles good English, primarily in comparison to other Latin-alphabet languages. These features are used in a separate maximum-entropy classifier whose output is a single feature to add to the MD classifier. Additional features, primarily to distinguish English from non-language input, are added to MD as well. An example is the minimum of the number of letters and the number of digits in the "word", which when greater than zero often indicates database detritus. Then we run the MD classifier enhanced with these new robustness-oriented features. We evaluate using a detection-error-trade-off (DET) (Martin et al., 1997) analysis, in addition to traditional precision/recall/F-measure.

This paper is organized as follows. Section 2 discusses previous work. Section 3 describes the baseline maximum-entropy-based MD system. Section 4 introduces enhancements to the system to achieve robustness. Section 5 describes databases used for experiments, which are discussed in Section 6, and Section 7 draws conclusions and plots future work.

2 Previous work on mention detection

The MD task has close ties to named-entity recognition, which has been the focus of much recent research (Bikel et al., 1997; Borthwick et al., 1998; Tjong Kim Sang, 2002; Florian et al., 2003; Benajiba et al., 2009), and has been at the center of several evaluations: MUC-6, MUC-7, CoNLL'02 and CoNLL'03 shared tasks. Usually, in computational-linguistics literature, a named entity represents an instance of either a location, a person, an organization, and the named-entity-recognition task consists of identifying each individual occurrence of names of such an entity appearing in the text. As stated earlier, in this paper we are interested in identification and classification of textual references to object/abstraction *mentions*, which can be either named, nominal or pronominal. This task has been a focus of interest in ACE since 2003. The recent ACE evaluation campaign was in 2008.

Effort to handle noisy data is still limited, especially for scenarios in which the system at decoding time does not have prior knowledge of the input data source. Previous work dealing with unstructured data assumes the knowledge of the input data source. As an example, E. Minkov *et al.* (Minkov et al., 2005) assume that the input data is text from e-mails, and define special features to enhance the detection of named entities. Miller *et al.* (Miller et al., 2000) assume that the input data is the output of a speech or optical character recognition system, and hence extract new features for better named-entity recognition. In a different research problem, L. Yi *et al.* eliminate the noisy text from the document before

performing data mining (Yi et al., 2003). Hence, they do not try to process noisy data; instead, they remove it. The approach we propose in this paper does not assume prior knowledge of the data source. Also we do not want to eliminate the noisy data, but rather attempt to detect the appropriate mentions, if any, that appear in that portion of the data.

3 Mention-detection algorithm

Similarly to classical NLP tasks such as base phrase chunking (Ramshaw and Marcus, 1999) and named-entity recognition (Tjong Kim Sang, 2002), we formulate the MD task as a sequence-classification problem, by assigning to each word token in the text a label indicating whether it starts a specific mention, is inside a specific mention, or is outside any mentions. We also assign to every non-outside label a class to specify entity type *e.g.* person, organization, location, etc. We are interested in a statistical approach that can easily be adapted for several languages and that has the ability to integrate easily and make effective use of diverse sources of information to achieve high system performance. This is because, similar to many NLP tasks, good performance has been shown to depend heavily on integrating many sources of information (Florian et al., 2004). We choose a Maximum Entropy Markov Model (MEMM) as described previously (Florian et al., 2004; Zitouni and Florian, 2009). The maximum-entropy model is trained using the *sequential conditional generalized iterative scaling* (SCGIS) technique (Goodman, 2002), and it uses a *Gaussian prior* for regularization (Chen and Rosenfeld, 2000)[1].

3.1 Mention detection: standard features

The featues used by our mention detection systems can be divided into the following categories:

1. **Lexical Features** Lexical features are implemented as token n-grams spanning the current token, both preceding and following it. For a token x_i, token n-gram features will contain the previous $n-1$ tokens ($x_{i-n+1}, \ldots x_{i-1}$) and the following $n-1$ tokens ($x_{i+1}, \ldots x_{i+n-1}$). Setting n equal to 3 turned out to be a good choice.

2. **Gazetteer-based Features** The gazetteer-based features we use are computed on tokens. The gazetteers consist of several class of dictionaries: including person names, country names, company names, etc. Dictionaries contain single names such as `John` or `Boston`, and also phrases such as `Barack Obama`, `New York City`, or `The United States`. During both training and decoding, when we encounter in the text a token or a sequence of tokens that completely matches an entry in a dictionary, we fire its corresponding class.

The use of this framework to build MD systems for clean English text has given very competitive results at ACE evaluations (Florian et al., 2006). Trying other classifiers is always a good experiment, which we didn't pursue here for two reasons: first, the MEMM system used here is state-of-the-art, as proven in evaluations and competitions – while it is entirely possible that another system might get better results, we don't think the difference would be large. Second, we are interested in ways of improving performance on noisy data, and we expect any system to observe similar degradation in performance when presented with unexpected input – showing results for multiple classifier types might very well dilute the message, so we stuck to one classifier type.

4 Enhancements for robustness

As stated above, our goal is to skip spans of characters which do not lend themselves to target-language MD, while minimizing impact on MD for target-language text, with English as the initial target language for our experiments. More specifically, our task is to process data automatically in any unpredetermined format from any source, during which we strive to avoid outputting spurious mentions on:

- non-language material, such as mark-up tags and other data-set detritus, as well as non-text data such as code or binaries likely mistakenly submitted to the MD system,
- non-target-character-set material, here, non-Latin-alphabet material, such as Arabic and Chinese in their native character sets, and
- target-character-set material not in the target language, here, Latin-alphabet languages other than English.

It is important to note that this is not merely a document-classification problem; this non-target data is often interspersed with valid input text.

[1] Note that the resulting model cannot really be called a maximum-entropy model, as it does not yield the model which has the maximum entropy (the second term in the product), but rather is a maximum-*a-posteriori* model.

Mark-up is the obvious example of interspersing; however, other categories of non-target data can also interleave tightly with valid input. A few examples:

- English text is sometimes infixed right in a Chinese sentence, such as 其他BBC网站
- some translation algorithms will leave unchanged an untranslatable word, or will transliterate it into the target language using a character convention which may not be a standard known to the MD engine, and
- some target-alphabet-but-non-target-language material will be compatible with the target language, particularly people's names. An example with English as the target language is `Barack Obama` in the Spanish text `...presidente de Estados Unidos, Barack Obama, dijo el da 24 que ....`

Therefore, to minimize needless loss of processable material, a robustness algorithm ideally does a sliding analysis, in which, character-by-character or word-by-word, material may be deemed to be suitable to process. Furthermore, a variety of strategies will be needed to contend with the diverse nature of non-target material and the patterns in which it will appear among valid input.

Accordingly, the following is a summary of algorithmic enhancements to MD:

1. detection of standard file formats, such as SGML, and associated detagging,
2. segmentation of the file into target- vs. non-target-character-set passages, such that the latter not be processed further,
3. tokenization to determine word and sentence units, and
4. MD, augmented as follows:
   - Sentence-level categorization of likelihood of good English.
   - If "clean" English was detected, run the same clean baseline model as described in Section 3.
   - If the text is determined to be a bad fit to English, run an alternate maximum-entropy model that is heavily based on gazetteers, using only context-independent (*e.g.* primarily gazetteer-based) features, to catch isolated obvious English/English-compatible names embedded in otherwise-foreign text.
   - If in between "clean" and "bad", use a "mixed" maximum-entropy MD model whose training data and feature set are augmented to handle interleaving of English with mark-up and other languages.

These MD-algorithm enhancements will be described in the following subsections.

4.1 Detection and detagging for standard file formats

Some types of mark-up are well-known standards, such as SGML (Warmer and van Egmond, 1989). Clearly the optimal way of dealing with them is to apply detectors of these specific formats, and associated detaggers, as done previously (Yi et al., 2003). For this reason, standard mark-up is not a subject of the current study; rather, our concern is with mark-up peculiar to specific data sets, as described above, and so while this step is part of our overall strategy, it is not employed in the present experiments.

4.2 Character-set segmentation

Some entity mentions may be recognizable in a non-target language which shares the target-language's character set, for example, a person's name recognizable by English speakers in an otherwise-not-understandable Spanish sentence. However, non-target character sets, such as Arabic and Chinese when processing English, represent pure noise for an IE system. Therefore, deterministic character-set segmentation is applied, to mark non-target-character-set passages for non-processing by the remainder of the system, or, in a multilingual system, to be diverted to a subsystem suited to process that character set. Characters which can be ambiguous with regard to character set, such as some punctuation marks, are attached to target-character-set passages when possible, but are not considered to break non-target-character-set passages surrounding them on both sides.

4.3 Tokenization

Subsequent processing is based on determination of the language of target-alphabet text. The fundamental unit of such processing is target-alphabet word, necessitating tokenization at this point into word-level units. This step includes punctuation sepa-

ration as well as the detction of sentence boundary (Zimmerman et al., 2006).

4.4 Robust mention detection

After preprocessing steps presented earlier, we detect mentions using a cascaded approach that combines several MD classifiers. Our goal is to select among maximum-entropy MD classifiers trained separately to represent different degrees of "noisiness" occurring in many genres of data, including machine-translation output, informal communications, mixed-language material, varied forms of non-standard database mark-up, etc. We somewhat-arbitrarily choose to employ three classifiers as described below. We select a classifier based on a sentence-level determination of the material's fit to the target language. First, we build an n-gram language model on clean target-language training text. This language model is used to compute the perplexity (PP) of each sentence during decoding. The PP indicates the quality of the text in the target-language (*i.e.* English) (Brown et al., 1992); the lower the PP, the cleaner the text. A sentence with a PP lower than a threshold θ_1 is considered "clean" and hence the "clean" baseline MD model described in Section 3 is used to detect mentions of this sentence. The clean MD model has access to standard features described in Section 3.1. In the case where a sentence looks particularly badly matched to the target language, defined as $PP > \theta_2$, we use a "*gazetteer-based*" model based on a dictionary look-up to detect mentions; we retreat to seeking known mentions in a context-independent manner reflecting that most of the context consists of out-of-vocabulary words. The gazetteer-based MD model has access only to gazetteer information and does not look to lexical context during decoding, reflecting the likelihood that in this poor material, words surrounding any recognizable mention are foreign and therefore unusable. In the case of an in-between determination, that is, a sentence with $\theta_1 < PP < \theta_2$, we use a "*mixed*" MD model, based on augmenting the training data set and the feature set as described in the next section. The values of θ_1 and θ_2 are estimated empirically on a separate development data set that is also used to tune the Gaussian prior (Chen and Rosenfeld, 2000). This set contains a mix of clean English and Latin-alphabet-but-non-English text that is not used for traning and evaluation.

The advantage of this combination strategy is that we do not need pre-defined knowledge of the text source in order to apply an appropriate model. The selection of the appropriate model to use for decoding is done automatically based on PP value of the sentence. We will show in the experiments section how this combination strategy is effective not only in maintaining good performance on a clean English text but also in improving performance on non-English data when compared to other source-specific MD models.

4.5 Mixed mention detection model

The mixed MD model is designed to process "sentences" mixing English with non-English, whether foreign-language or non-language material. Our approach is to augment model training compared to the clean baseline by adding non-English, mixed-language, and non-language material, and to augment the model's feature set with language-identification features more localized than the sentence-level perplexity described above, as well as other features designed primarily to distinguish non-language material such as mark-up codes.

4.5.1 Language-identification features

We apply an n-gram-based language classifier (Prager, 1999) to variable-length sliding windows as follows. For each word, we run 1- through 6-preceding-word windows through the classifier, and 1- through 6-word windows beginning with the word, for a total of 12 windows, yielding for each window a result like:

```
0.235 Swedish
0.148 English
0.134 French
...
```

For each of the 12 results, we extract three features: the identity of the top-scoring language, here, Swedish; the confidence score in the top-scoring language, here, 0.235; and the score difference between the target language (English for these experiments) and the top-scoring non-target language, here, $0.148 - 0.235 = -0.087$. Thus we have a 36-feature vector for each word. We bin these and use them as input to a maximum-entropy classifier (separate from the MD classifier) which outputs "English" or "Non-English", and a confidence score. These scores in turn are binned into six categories to serve as a "how-English-is-it" feature in the augmented MD model. The language-identification classifier and the maximum-entropy "how-English" classifier are each trained on text data separate from

each other and from the training and test sets for MD.

4.5.2 Additional features

The following features are designed to capture evidence of whether a "word" is in fact linguistic material or not: number of alphabetic characters, number of characters, maximum consecutive repetitions of a character, numbers of non-alphabetic and non-alphanumeric characters, fraction of characters which are alphabetic, fraction alphanumeric, and number of vowels. These features are part of the augmentation of the mixed MD model relative to the clean MD model.

5 Data sets

Four data sets are used for our initial experiments. One, "English", consists of 367 documents totaling 170,000 words, drawn from web news stories from various sources and detagged to be plain text. This set is divided into 340 documents as a training set and 27 for testing, annotated as described in more detail elsewhere (Han, 2010). These data average approximately 21 annotated mentions per 100 words.

The second set, "Latin", consists of 23 detagged web news articles from 11 non-English Latin-alphabet languages totaling 31,000 words. Of these articles, 12 articles containing 19,000 words are used as a training set, with the remaining used for testing, and each set containing all 11 languages. They are annotated using the same annotation conventions as "English", and from the perspective of English; that is, only mentions which would be clear to an English speaker are labeled, such as `Barack Obama` in the Spanish example in Section 4. For this reason, these data average only approximately 5 mentions per 100 words.

The third, "Transactions", consists of approximately 60,000 words drawn from a text data set logging real financial transactions. Figure 1 shows example passages from this database, anonymized while preserving the character of the content.

This data set logs transactions by a staff of customer-service representatives. English is the primary language, but owing to international clientele, occasionally representatives communicate in other languages, such as the German here, or in English but mentioning institutions in other countries, here, a Czech bank. Interspersed among text are codes specific to this application which delineate and identify various information fields and punctuate long passages. The application also places constraints on legal characters, leading to the unusual representation of underline and the "at" sign as shown, making for an e-mail address which is human-readable but likely not obvious to a machine. Abbreviations represent terms particularly common in this application area, though they may not be obvious without adapting to the application; these include standards like `HUF`, a currency code which stands for Hungarian forint, and financial-transaction peculiarities like `BNF` for "beneficiary" as seen in Figure 1. In short, good English is interspersed with non-language content, foreign-language text, and rough English like data-entry errors and haphazard abbreviations. These data average 4 mentions per 100 words.

Data sets with peculiarities analogous to those in this Transactions set are commonplace in a variety of settings. Training specific to data sets like this is often infeasible due to lack of labeled data, insufficient data for training, and the multitude of such data formats. For this reason, we do not train on Transactions, letting our testing on this data set serve as an example of testing on such data formats unseen.

6 Experiments

MD systems were trained to recognize the 116 entity-mention types shown in Table 1, annotated as described previously (Han, 2010). The clean-data classifier was trained on the English training data using the feature set described in Section 3.1. The classifier for "mixed"-quality data and the "gazetteer" model were each trained on that set plus the "Latin" training set and the supplemental set. In addition, "mixed" training included the additional features described in Section 4.5. The framework used to build the baseline MD system is similar to the one we used in the ACE evaluation[2]. This system has achieved competitive results with an F-measure of 82.7 when trained on the seven main types of ACE data with access to wordnet and part-of-speech-tag information as well as output of other MD and named-entity recognizers (Zitouni and Florian, 2008).

It is instructive to evaluate on the individual component systems as well as the combination, despite the fact that the individual components are not well-suited to all the data sets, for example, the mixed and gazetteer systems being a poorer fit to the English task than the baseline, and vice versa for the

[2]NIST's ACE evaluation plan: http://www.nist.gov/speech/tests/ace/index.htm

age	event-custody	facility	people	date
animal	event-demonstration	food	percent	duration
award	event-disaster	geological-object	person	e-mail-address
cardinal	event-legal	geo-political	product	measure
disease	event-meeting	law	substance	money
event	event-performance	location	title-of-a-work	phone-number
event-award	event-personnel	ordinal	vehicle	ticker-symbol
event-communication	event-sports	organ	weapon	time
event-crime	event-violence	organization	web-address	

Table 1: Entity-type categories used in these experiments. The eight in the right-most column are not further distinguished by mention type, while the remaining 36 are further classified as named, nominal or pronominal, for a total of $36 \times 3 + 8 = 116$ mention labels.

	English			**Latin**			**Transactions**		
	P	R	F	P	R	F	P	R	F
Clean	78.7	73.6	76.1	16.0	40.0	22.9	19.5	32.2	24.3
Mixed	77.9	69.7	73.6	78.5	55.9	65.3	37.1	47.8	41.7
Gazetteer	76.9	66.2	71.1	77.8	55.5	64.8	36.5	47.5	41.3
Combination	78.1	73.2	**75.6**	80.4	56.0	**66.0**	38.5	49.1	**43.2**

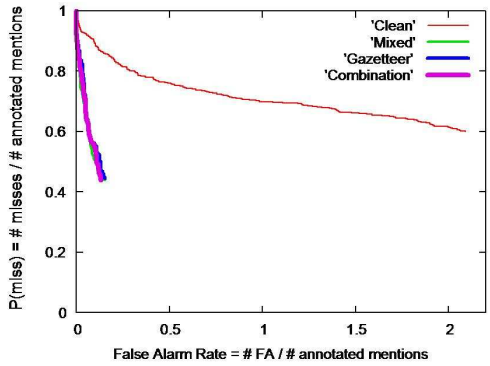
Table 2: Performance of clean, mixed, and gazetteer-based mention detection systems as well as their combination. Performance is presented in terms of Precision (P), Recall (R), and F-measure (F).

non-target data sets. Precision/recall/F-measure results are shown in Table 2. Not surprisingly, the baseline system, intended for clean data, performs poorly on noisy data. The mixed and gazetteer systems, having a variety of noisy data in their training set, perform much better on the noisy conditions, particularly on Latin-alphabet-non-English data because that is one of the conditions included in its training, while Transactions remains a condition not covered in the training set and so shows less improvement. However, because the mixed classifier, and moreso the gazetteer classifier, are oriented to noisy data, on clean data they suffer in performance by 2.5 and 5 F-measure points, respectively. But system combination serves us well: it recovers all but 0.5 F-measure point of this loss, while also actually performing better on the noisy data sets than the two classifiers specifically targeted toward them, as can be seen in Table 2. It is important to note that the major advantage of using the combination model is the fact that we do not have to know the data source in order to select the appropriate MD model to use. We assume that the data source is unknown, which is our claim in this work, and we show that we obtain better performance than using source-specific MD models. This reflects the fact that a noisy data set will in fact have portions with varying degrees of "noise", so the combination outperforms any single model targeted to a single particular level of noise, enabling the system to contend with such variability without the need for presegregating sub-types of data for noise level. The obtained improvement from the system combination over all other models is statistically significant based on the stratified bootstrap re-sampling significance test (Noreen, 1989). We consider results statistically significant when $p < 0.05$, which is the case in this paper. This approach was used in the named-entity-recognition shared task of CoNNL-2002[3].

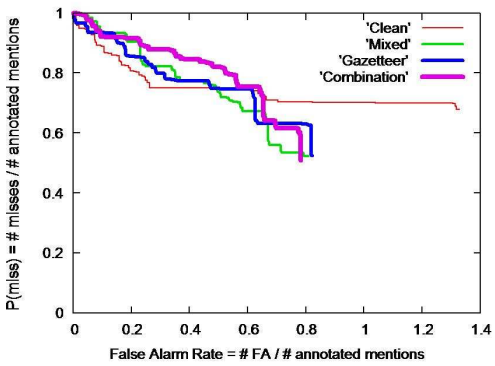
It should be noted that some completely-non-target types of data, such as non-target-character set data, have been omitted from analysis here. Including them would make our system look comparatively stronger, as they would have only spurious mentions and so generate false alarms but no correct mentions in the baseline system, while our system deterministically removes them.

As mentioned above, we view MD robustness primarily as an effort to eliminate, relative to a baseline system, large volumes of spurious "mentions" detected in non-target input content, while minimiz-

[3]http://www.cnts.ua.ac.be/conll2002/ner/



(a) DET plot for clean (baseline), mixed, gazetteer, and combination MD systems on the Latin-alphabet-non-English text. The clean system (upper curve) performs far worse than the other three systems designed to provide robustness; these systems in turn perform nearly indistinguishably.



(b) DET plot for clean (baseline), mixed, gazetteer, and combination MD systems on the Transactions data set. The clean system (upper/longer curve) reaches far higher false-alarm rates, while never approaching the lower miss rates achievable by any of the other three systems, which in turn perform comparably to each other.

Figure 2: DET plots for Latin-alphabet-non-English and Transactions data sets

ing disruption of detection in target input. A secondary goal is recall in the event of occasional valid mentions in such non-target material. Thus, as input material degrades, precision increases in importance relative to recall. As such, we view precision and recall asymmetrically on this task, and so rather than evaluating purely in terms of F-measure, we perform a detection-error-trade-off (DET) (Martin et al., 1997) analysis, in which we plot a curve of miss rate on valid mentions vs. false-alarm rate, with the curve traced by varying a confidence threshold across its range. We measure false-alarm and miss rates relative to the number of actual mentions annotated in the data set:

$$\text{FA rate} = \frac{\#\text{ false alarms}}{\#\text{ annotated mentions}} \quad (1)$$

$$\text{Miss rate} = \frac{\#\text{ misses}}{\#\text{ annotated mentions}} \quad (2)$$

where false alarms are "mentions" output by the system but not appearing in annotation, while misses are mentions which are annotated but do not appear in the system output. Each mention is treated equally in this analysis, so frequently-recurring entity/mention types weigh on the results accordingly.

Figure 2a shows a DET plot for the clean, mixed, gazetteer, and combination systems on the "Latin" data set, while Figure 2b shows the analogous plot for the "Transactions" data set. The drastic gains made over the baseline system by the three experimental systems are evident in the plots. For example, on Latin, choosing an operating point of a miss rate of 0.6 (nearly the best achievable by the clean system), we find that the robustness-oriented systems eliminate 97% of the false alarms of the clean baseline system, as the plot shows false-alarm rates near 0.07 compared to the baseline's of 2.08. Gains on Transaction data are more modest, owing to this case representing a data genre not included in training. It should be noted that the jaggedness of the Transaction curves traces to the repetitive nature of some of the terms in this data set.

In making a system more oriented toward robustness in the face of non-target inputs, it is important to quantify the effect of these systems being less-oriented toward clean, target-language text. Figure 3 shows the analogous DET plot for the English test set, showing that achieving robustness through the combination system comes at a small cost to accuracy on the text the original system is trained to process.

7 Conclusions

For information-extraction systems to be useful, their performance must degrade gracefully when confronted with inputs which deviate from ideal and/or derive from unknown sources in unknown formats. Imperfectly-translated, mixed-language, marked-up text and non-language material must not

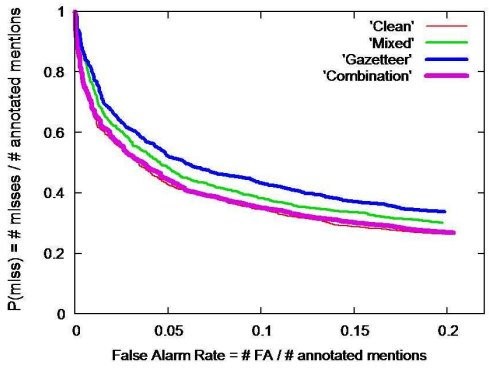


Figure 3: DET plot for clean (baseline), mixed, gazetteer, and combination MD systems on clean English text, verifying that performance by the clean system (lowest curve) is very closely approximated by the combination system (second-lowest curve), while the mixed system performs somewhat worse and the gazetteer system (top curve), worse still, reflecting that these systems are increasingly oriented toward noisy inputs.

be processed in a garbage-in-garbage-out fashion merely because the system was designed only to handle clean text in one language. Thus we have embarked on information-extraction-robustness work, to improve performance on imperfect inputs while minimizing disruption of processing of clean text. We have demonstrated that for one IE task, mention detection, a multi-faceted approach, motivated by the diversity of input data imperfections, can eliminate a large proportion of the spurious outputs compared to a system trained on the target input, at a relatively small cost of accuracy on that target input. This outcome is achieved by a system-combination approach in which a perplexity-based measure of how well the input matches the target language is used to select among models designed to deal with such varying levels of noise. Rather than relying on explicit recognition of genre of source data, the experimental system merely does its own assessment of how much each sentence-sized chunk matches the target language, an important feature in the case of unknown text sources.

Chief among directions for further work is to continue to improve performance on noisy data, and to strengthen our findings via larger data sets. Additionally, we look forward to expanding analysis to different types of imperfect input, such as machine-translation output, different types of mark-up, and different genres of real data. Further work should also explore the degree to which the approach to achieving robustness must vary according to the target language. Finally, robustness work should be expanded to other information-extraction tasks.

Acknowledgements

The authors thank Ben Han, Anuska Renta, Veronique Baloup-Kovalenko and Owais Akhtar for their help with annotation. This work was supported in part by DARPA under contract HR0011-08-C-0110.

References

Y. Benajiba, M. Diab, and P. Rosso. 2009. Arabic named entity recognition: A feature-driven study. *In the special issue on Processing Morphologically Rich Languages of the IEEE Transaction on Audio, Speech and Language*.

D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP-97*, pages 194–201.

A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. C. Lai, and R. L. Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1), March.

S. Chen and R. Rosenfeld. 2000. A survey of smoothing techniques for ME models. *IEEE Transaction on Speech and Audio Processing*.

R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. 2003. Named entity recognition through classifier combination. In *Conference on Computational Natural Language Learning - CoNLL-2003*, Edmonton, Canada, May.

R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N Nicolov, and S Roukos. 2004. A statistical model for multilingual entity detection and tracking. In *Proceedings of HLT-NAACL 2004*, pages 1–8.

R. Florian, H. Jing, N. Kambhatla, and I. Zitouni. 2006. Factorizing complex models: A case study in mention detection. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 473–480, Sydney, Australia, July. Association for Computational Linguistics.

J. Goodman. 2002. Sequential conditional generalized iterative scaling. In *Proceedings of ACL'02*.

D. B. Han. 2010. Klue annotation guidelines - version 2.0. Technical Report RC25042, IBM Research, August.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. 1997. The DET curve in assessment of detection task performance. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1895–1898. Rhodes, Greece.

D. Miller, S. Boisen, R. Schwartz, R. Stone, and R. Weischedel. 2000. Named entity extraction from noisy input: speech and OCR. In *Proceedings of the sixth conference on Applied natural language processing*, pages 316–324, Morristown, NJ, USA. Association for Computational Linguistics.

E. Minkov, R. C. Wang, and W. W. Cohen. 2005. Extracting personal names from email: Applying named entity recognition to informal text. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 443–450, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

E. W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley Sons.

J. F. Pitrelli, B. L. Lewis, E. A. Epstein, M. Franz, D. Kiecza, J. L. Quinn, G. Ramaswamy, A. Srivastava, and P. Virga. 2008. Aggregating Distributed STT, MT, and Information Extraction Engines: The GALE Interoperability-Demo System. In *Interspeech*. Brisbane, NSW, Australia.

J. M. Prager. 1999. Linguini: Language identification for multilingual documents. In *Journal of Management Information Systems*, pages 1–11.

L. Ramshaw and M. Marcus. 1999. Text chunking using transformation-based learning. In S. Armstrong, K.W. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 157–176. Kluwer.

E. F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

J. Warmer and S. van Egmond. 1989. The implementation of the Amsterdam SGML parser. *Electron. Publ. Origin. Dissem. Des.*, 2(2):65–90.

L. Yi, B. Liu, and X. Li. 2003. Eliminating noisy information in web pages for data mining. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 296–305, New York, NY, USA. ACM.

M. Zimmerman, D. Hakkani-Tur, J. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu. 2006. The ICSI+ multilingual sentence segmentation system. In *Interspeech*, pages 117–120, Pittsburgh, Pennsylvania, September.

I. Zitouni and R. Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of EMNLP'08*, Honolulu, Hawaii, October.

I. Zitouni and R. Florian. 2009. Cross-language information propagation for Arabic mention detection. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–21.