Increasing NER recall with minimal precision loss

Jasper Kuperus
Sogeti Nederland B.V.
Vianen, The Netherlands
Email: mail@jasperkuperus.nl

Cor J. Veenman
Netherlands Forensic Institute, Kecida
The Hague, The Netherlands
Email: c.veenman@nfi.minvenj.nl

Maurice van Keulen
University of Twente, EEMCS
Enschede, The Netherlands
Email: m.vankeulen@utwente.nl

***Abstract*—Named Entity Recognition (NER) is broadly used as a first step toward the interpretation of text documents. However, for many applications, such as forensic investigation, recall is currently inadequate, leading to loss of potentially important information. Entity class ambiguity cannot be resolved reliably due to the lack of context information or the exploitation thereof. Consequently, entity classification introduces too many errors, leading to severe omissions in answers to forensic queries.**

We propose a technique based on *multiple candidate labels*, effectively postponing decisions for entity classification to query time. Entity resolution exploits user feedback: a user is only asked for feedback on entities relevant to his/her query. Moreover, giving feedback can be stopped anytime when query results are considered good enough. We propose several interaction strategies that obtain increased recall with little loss in precision.

I. Introduction

Named Entity Recognition (NER) is a useful technique in many different domains. Although NER has already been researched intensely, it remains a difficult task [1]. One of the main challenges in NER is the ambiguity of the recognized named entities [2]. Examples of ambiguity are whether a numerical value represents a phone number or a social security number (semantic ambiguity), or what the correct boundaries for an entity are (structural ambiguity, e.g., `Lake Como` or `Como`). Such ambiguity is often trivial for a human with his cognitive abilities easily interpreting the context [3], but extremely difficult for an automated process.

With the application of forensic investigation in mind, a typical use is network analysis — finding among others relations between people, companies, and addresses, a use well-supported by linking entities occurring in case data. Here, recall is more important than precision: we do not want to lose potentially important information, such as the possibility of a relationship between two persons, which need to be manually confirmed anyway, hence some noise is acceptable.

Most NER solutions identify multiple alternative interpretations of an entity. Machine learning approaches typically assign probabilities to these alternatives, like e.g. Conditional Random Fields (CRFs) [4], [5]. However, most approaches decide on the most probable alternative as the correct classification. For example, a phrase estimated to be either a company or a person name with likelihoods 0.6 and 0.4 respectively does not make *company* the absolute correct answer, introducing the risk of assigning the wrong type to an entity. These decisions will inevitably result in errors and therefore lower recall.

A method to prevent these extraction errors is to work with multiple *candidate* classifications, effectively postponing the decision or not making a decision at all. Every identified candidate can be assigned a confidence score that correlates with the probability of this candidate being the correct one [6]. We call *Probabilistic Named Entity Recognition (PNER)* the approach to keep *all* possible candidates with their probability.

PNER results in higher recall but also in uncertainty in the data, i.e., lower precision. A method to reduce uncertainty and raise precision is by asking the user to resolve ambiguous situations. We call *Targeted Feedback* the process of finding the query that results in the highest gain in overall confidence, therefore making user feedback more effective, striving towards a minimal need for user feedback.

Contributions The main contributions of this paper are

- PNER approach enabling postponing and eliminating entity classification decisions, and
- Targeted Feedback strategies for quickly raising data quality with minimal user effort.

⇒ Fewer omissions in answers to (forensic) queries with control over trade-off between user effort and precision.

We first explain our PNER approach in Section II. We then present multiple strategies for Targeted Feedback in Section III that are experimentally compared in Sections IV and V. We end with a discussion of future work and conclusions.

II. Probabilistic NER

A. Ambiguity

There are different kinds of ambiguity that play a role in NER. Van Keulen et al. [3] define three kinds of ambiguity:

- **Semantic ambiguity** refers to the classification of an entity, does Paris refer to a name or a location?
- **Structural ambiguity** refers to ambiguity regarding the structure and boundaries of entities, e.g., is the word `Lake` part of the entity for the location `Lake Como`?
- **Reference ambiguity** refers to the question to which real world object an entity refers, e.g., does the location `Paris` refer to the Paris in France or one of the other 158 Paris instances found in GeoNames [7]?

The problem of reference ambiguity is actively researched, e.g., [8], [9], [10], and our work is complementary to this research.

978-0-7695-5062-6/13 $26.00 © 2013 IEEE
DOI 10.1109/EISIC.2013.23

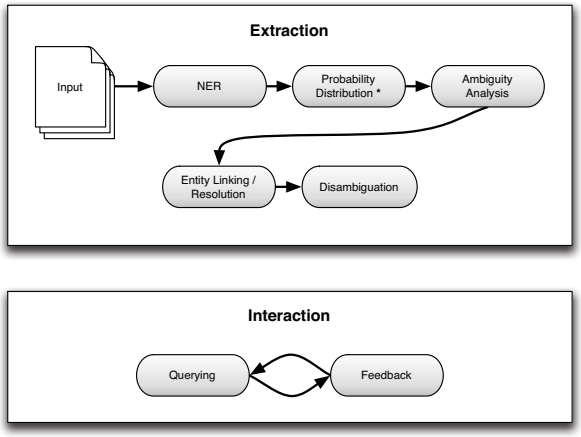

Fig. 1. The "Extraction" and "Interaction" subprocesses of PNER.

B. PNER process

Since we propose to postpone decisions to query time, we distinguish the subprocesses "Extraction" and "Interaction" (see Figure 1).

1) Extraction subprocess: The extraction subprocess can start in a traditional NER manner. One either adapts an existing NER technique to produce multiple candidates with probabilities or defines a subsequent probability distribution step. The candidate set is first analyzed for structural and semantic ambiguity. Subsequently, the candidates are linked to entries from a knowledge base introducing reference ambiguity. In the remainder of the paper, we focus on semantic and structural ambiguity only. The extraction subprocess ends with a disambiguation step. Although automatic disambiguation can introduce the very errors PNER attempts to avoid, some filtering of obvious cases and redistribution of probabilities given reference knowledge is beneficial.

2) Interaction subprocess: Using probabilistic database technology, the uncertain data produced in extraction can be queried immediately resulting initially in a low quality query result with many alternatives. Targeted Feedback constructs and poses a question to the user whose answer reduces the likelihood or completely eliminates NER candidates, which in turn improve the query result. This process is repeated until the user is satisfied with its quality.

Resolving all ambiguity in a big dataset through user feedback is a labor-intensive task and not the goal of Targeted Feedback. In some situations, resolving ambiguity by 50% might just be enough for a user to confidently answer his/her question. Also, Targeted Feedback stays within the scope of the query, hence avoiding solving ambiguity concerning entities not relevant to the current information need.

C. Adapting existing NER techiques for PNER

As mentioned, it is possible to adapt some existing NER techniques to produce multiple candidates with probabilities. Although performing hard entity classification, often the absolute decisions are internally based on a probabilistic approach. For example, the Stanford CRF Classifier [11], [12], which we use in our experiments, provides information on the probabilities underlying its classifications.

```
Brussels O=0.005  ORG=0.111  MISC=0.012  PER=0.137  LOC=0.701
as       O=0.984  ORG=7.3E-6 MISC=0.015  PER=4.3E-6 LOC=2.4E-5
the      O=0.985  ORG=3.1E-5 MISC=0.014  PER=6.3E-6 LOC=2.9E-5
Belgian  O=0.998  ORG=6.2E-4 MISC=8.6E-6 PER=3.5E-4 LOC=8.4E-4
and      O=0.852  ORG=2.5E-5 MISC=0.147  PER=2.4E-6 LOC=1.8E-4
European O=1.0E-4 ORG=0.315  MISC=0.192  PER=0.212  LOC=0.206
Centre   O=5.8E-4 ORG=0.301  MISC=0.178  PER=0.102  LOC=0.217
Brussels O=1.1E-4 ORG=0.202  MISC=0.163  PER=0.203  LOC=0.331
...
```

Fig. 2. Stanford CRF Classifier Classification Result Format

As can be seen in Figure 2, it determines per token a probability for each class. These can be directly taken as candidates (semantic ambiguity). A threshold of 0.01 is applied to eliminate marginal candidates. Candidates for structural ambiguity are constructed by detecting subsequent tokens with the same type that combined have a non-marginal probability. The probability of such a multi-token entity can be calculated using Stanford's support for obtaining conditional probabilities and the *multiplication axiom* [13].

$$P(A_1 A_2 \dots A_n) = \prod_{i=1}^{n} P(A_i \mid A_1 \dots A_{i-1}) \qquad (1)$$

For example, $P(\texttt{ European Centre=ORG }) = P(\texttt{ European=ORG }) \cdot P(\texttt{ Centre=ORG } \mid \texttt{ European=ORG })$.

D. Probabilistic data model

The PNER results are label *candidates* with probabilities. Such results can be readily stored in a probabilistic database. We developed a simple *relational model* for the experiments in Section IV.

Semantic and structural ambiguity is local, i.e., candidates come in *bundles* covering a few consecutive words, e.g., the `European Centre Brussels` we saw earlier. The relational model is based on two tables: Entities and Classifications. The former holds for each entity candidate its id, entity name, boundary (begin and end position), parent boundary (boundary of entity bundle) and structural probability. The latter holds for each entity candidate and each of its possible classes, its id, class, and semantic probability. A candidate for the example bundle is $\langle 42, \texttt{European Centre}, (0,15), (0,24), 0.2\rangle$, which could, e.g., have two possible classes: $\langle 42, \texttt{LOC}, 0.3\rangle$ and $\langle 42, \texttt{ORG}, 0.5\rangle$.

This relational model cannot express mutual exclusiveness. For example, it assigns non-zero probabilities to impossible worlds such as $\texttt{European} \wedge \texttt{European Centre Brussels}$. The more complex random variables model solves this problem [14].

III. Targeted Feedback

The extraction subprocess of PNER is fully automatic and, in constrast to regular NER approaches, it retains all possible interpretations. Even though the system may not be able to

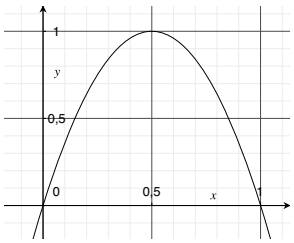

Fig. 3. Graph of MAEF $AmbScore$ function (Equation 2)

provide the correct interpretation with the highest probability, it is careful in not missing the correct interpretation. It will be among all the possibilities, i.e., recall is high. Accordingly, an explosion of possibilities is likely. By employing probabilistic database techniques, this volume of uncertain data can be handled. Clearly, query results will exhibit quite some noise from all these possibilities, i.e., precision is low. If the user is unsatisfied with the precision of the results, he can raise it by giving feedback on ambiguous situations relevant to the query. Feedback is given by answering questions posed by the system that resolve some of the ambiguity. Note that the system does not take automated decisions, the user decides through feedback.

User effort in answering questions is costly and not every feedback raises precision with the same amount. Therefore, user effort can be reduced if the system manages to pose those questions first whose answer would increase precision most. We call this *Targeted* Feedback. Multiple strategies can be defined for choosing which question to pose next.

Besides these Question Proposal Strategies, we also distinguish Answer Handling Strategies. These vary in what information they derive from the answer, for example, by exploiting opportunities for extrapolation. In the end, Targeted Feedback needs a combination of one of each strategies.

In this paper, we experiment with several baseline and more refined strategies. Ideas for more elaborate and intelligent strategies are discussed in Section VI. The performance of a strategy is defined by both the speed of increase in precision and the moment of the biggest increase.

A. Question Proposal Strategies

We experiment with the following strategies:

1) Sequential: Baseline strategy that proposes to solve all ambiguous entities in the query answer in the order of their extraction from the original documents.

2) Random: Baseline strategy that randomly picks an ambiguous entity in the query answer.

3) Largest Bundle First (LBF): Structural ambiguity abounds in PNER resulting in many entity bundles. The LBF strategy poses the entity bundle with the most entities to the user first. The intuition is that solving structural and semantic ambiguity of many entities in one question is likely to give a high increase in precision.

4) Most Ambiguous Entity First (MAEF): LBF ignores probabilities. An entity with two candidate classifications A (0.95) and B (0.05) can be considered not very ambiguous. Similarly, a distribution A (0.9), B (0.07), C (0.03) can be considered much less ambiguous than A (0.44), B (0.32), C (0.24).[1] The intuition behind MAEF is to pose the entity first that has the highest ambiguity, and if this most ambiguous entity is part of an entity bundle, pose the entire bundle. We measure ambiguity with an *ambiguity score* function which is a quadratic function with its maximum at 0.5 and zero-crossings at 0 and 1 ($cl.prob$ is the probability of candidate classification cl for entity e).

$$AmbScore(e) \quad = \quad \sum_{cl \in e} -4(cl.prob - 0.5)^2 + 1 \qquad (2)$$

5) Most Ambiguous Cluster First (MACF): There are Answer Handling Strategies, such as *Statistical* (see below), that attempt to extrapolate answers of the user. For example, if the feedback on `Amsterdam` is very often `LOC`, then the probability of `Amsterdam` being `LOC` in all unresolved entities can be increased. Such an Answer Handling Strategy benefits from being combined with a Question Proposal Strategy that prefers entities containing often occurring entities.

The MACF strategy clusters entities by their name. $AmbScore(c) = \sum_{e \in c} AmbScore(e)$. From the cluster c with the highest score, the entity with the highest score is posed to the user, or the entity bundle if it is part of a bundle.

B. Answer Handling Strategies

Answer Handling Strategies define what information is derived from the feedback of the user, hence which actions are taken. Possible actions are elimination of candidates or adjusting and redistributing probabilities.

1) BasicDB: This strategy eliminates those entity candidates indicated to be incorrect by the user and assigns a probability of 1.0 to those indicated to be correct. The feedback can also be a new entity or classification, which is inserted into the database with probability 1.0. All other strategies should strictly extend this one, i.e., not refrain from taking these actions.

2) Statistical: This strategy attempts to extrapolate feedback on entities to others with the same name. It keeps track of a counter $ctr_{e,cl}$ per entity-classification combination, for example, $(\texttt{Amsterdam}, \texttt{LOC})$. Probabilities of all candidates with the same name are redistributed: $cl.prob{:=}cl.prob{\cdot}(ctr_{e,cl}{+}1)$ followed by a normalization of $cl.prob$ for each e. For example, a $60\%/20\%/20\%$ distribution for `Amsterdam` being `LOC`/`ORG`/`MISC` is redistributed to $75\%/12.5\%/12.5\%$ upon the first feedback that some other `Amsterdam` is a location.

IV. Experimental Setup

A. Data, Queries, and Strategies

We used the SoNaR [15], [16], [17], [18] corpus, a major reference corpus for contemporary written Dutch (Dec 2011).

[1]Note that one of A, B, or C can represent the class representing the case 'not an entity at all'.

Query	Description
A: *	All entities
B: *=PER	All persons
C: *=EVE	All events
D: *=LOC &2 *Dylan*	All locations within 2 lines of an entity containing 'Dylan'
E: *=PER &1 USA	All persons within 1 line of 'USA' entities

TABLE I
VALIDATION QUERIES

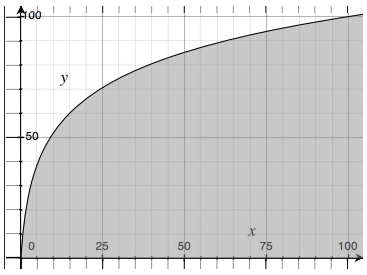

Fig. 4. Strategy Comparison Measure - F-measure Area

More specifically, we use the manually annotated wikipedia articles, partitioned into a training, development, and validation set, counting 96.450, 75.009 and 75.003 words respectively.

The validation queries in Table I are selected by browsing and executing queries on the development partition with the aim to find queries that vary result size, data quality, and type.

All combinations of the mentioned Question Proposal Strategies and Answer Handling Strategies are evaluated.

B. User Feedback Simulation

In the experiments, the activity of the user giving feedback is simulated based on the ground truth from SoNaR. Given the entity bundle proposed by the Question Proposal Strategy, the simulated user performs these steps:

1) Obtain ground truth for the classification of the entities and the boundaries of the bundle.
2) Approve correct entity/classification pairs (true positive).
3) For an entity whose correct classification is not among the candidates, add it (false negative).
4) For an entity in the ground truth that overlaps with the bundle but does not exist in the extracted bundle, add it (false negative).
5) Remove incorrect entity/classification pairs (false positive).

These simulator steps aim to closely resemble how a real user would provide feedback by (dis)approving candidates or adding missing entities or classifications.

C. Validation measures

Standard validation measures for NER are [19], [20]

$$Precision = \frac{\textit{\#correct entities found}}{\textit{\#total entities found}} \quad (3)$$

$$Recall = \frac{\textit{\#correct entities found}}{\textit{\#total correct entities}} \quad (4)$$

$$F{-}measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

These measures are developed for 'absolute' results and do not account for multiple alternatives nor probabilities. They are still useful for comparison with traditional NER, because a traditional NER result can be derived from PNER's probabilistic result by taking only the top most probable candidates and boundaries. Van Keulen et al. [21] define variants of $Precision$ and $Recall$ for probabilistic data as the expected value of $Precision$ and $Recall$ over all possible worlds. They proved that this is equivalent with replacing 'counting' in the traditional $Precision$ and $Recall$ formulas with a sum of probabilities. Intuitively, it 'counts' an answer by its probability.

We define a third $Recall$ measure, called $E_{100}(Recall)$, which does not take into account the probabilities. In this way, it determines the 'coverage', i.e., how many of the correct answers are covered by the candidates in the probabilistic answer, for a large part evaluating a correct setting of a cut-off threshold. Note that although the formula looks the same as for traditional $Recall$, the numbers of entities here refer to the entire candidate set.

$$E(Precision) = \frac{f(\textit{\#correct entities found})}{f(\textit{\#total entities found})} \quad (6)$$

$$E(Recall) = \frac{f(\textit{\#correct entities found})}{\textit{\#total correct entities}} \quad (7)$$

$$E_{100}(Recall) = \frac{\textit{\#correct entities found}}{\textit{\#total correct entities}} \quad (8)$$

$$f(x) = \sum_{(e,cl)\in x} (e.prob \cdot cl.prob) \quad (9)$$

Note that for a non-probabilistic result, i.e., without alternatives and a probability of 1 for each result, $E(Precision) = Precision$ and $E(Recall) = E_{100}(Recall) = Recall$. Furthermore, it is necessary to consider all three measures together as each can be trivially maximized individually judging a different aspect of result quality.

Finally, the performance of a strategy is defined by both the speed of the data quality increase and the moment in time of the biggest increase. No measure yet exists for this purpose, therefore we introduce the $F{-}measure\ Area$ defined as the area under the $F{-}measure$ graph, as illustrated in Figure 4. We apply this measure to the graph which has the number of questions n on the x-axis and the $F{-}measure$ on the y-axis. The area under the F-measure graph can be calculated using a Riemann sum:

$$F{-}measure\ Area = \frac{100}{n} \sum_{i=0}^{n} F{-}measure_i \quad (10)$$

V. RESULTS

A. Initial Query Performance

Comparing $Precision$ and $E(Precision)$ in Table II shows a higher value for $E(Precision)$ for all queries. This tells

Query	$Precision$	$E(Precision)$	$Recall$	$E(Recall)$	$E_{100}(Recall)$
A	43.13%	61.98%	52.17%	38.78%	87.33%
B	19.17%	28.20%	54.80%	40.35%	95.59%
C	5.56%	14.41%	39.76%	30.47%	65.06%
D	8.72%	15.85%	81.82%	69.56%	95.45%
E	19.79%	21.33%	90.48%	62.33%	95.24%

TABLE II
VALIDATION QUERIES - PERFORMANCE

Query	Precision	Recall
A	61.61%	57.60%
B	64.72%	66.51%
C	88.89%	38.55%
D	31.36%	84.09%
E	3.19%	95.24%

TABLE III
VALIDATION QUERIES - STANFORD NER

us that a considerable amount of *correct answers* in the dataset are not assigned the highest probability among their alternatives and are therefore not extracted with traditional NER. Comparing $Recall$ with $E_{100}(Recall)$ shows that PNER manages to extract much more of the correct answers among its alternatives, correct answers that would be missed with the absolute decision making of traditional NER.

The initial $Precision$ and $E(Precision)$ values seem quite low. However, Kuperus [14] provides an extensive explanation on how these values relate to the different kinds of ambiguity and how this results in the shown $Precision$ values.

B. Traditional NER

Our PNER approach introduces an additional dimension to the probabilistic results of Stanford NER [11], structural ambiguity. Therefore, Precision and Recall in Table II do not reflect traditional NER. Table III shows the performance of Stanford as-is. Comparing the $Recall$ of Stanford with $E_{100}(Recall)$ shows that PNER has a significantly better coverage of the correct answers than traditional NER, adding up to a difference over **29%** for queries **A** and **B**.

C. Strategy Performance

We have observed the general trend that for almost all combinations of strategies and queries both $Precision$ and $Recall$ approach 100% (not shown). This is due to the high initial $E_{100}(Recall)$ and the fact that a user can introduce new entities and classifications.

Table IV presents a performance comparison for all strategy combinations. The baseline strategy Sequential is outperformed in almost all cases. The other baseline strategy Random proved harder to beat.

Although the MACF strategy is introduced mainly as enhancement of MAEF and meant to be combined with Statistical, it is in most cases outperformed by both LBF and MAEF. The reason for this can be found in the fact that the query

		Queries					
		A	**B**	**C**	**D**	**E**	**Avg**
BasicDB	Sequential	69.93%	53.39%	29.22%	36.90%	**57.98%**	49.48%
	Random	79.12%	61.79%	32.21%	38.87%	57.17%	53.83%
	LBF	***81.08%***	***63.09%***	***32.94%***	*41.77%*	*57.49%*	***55.27%***
	MAEF	80.88%	62.03%	32.42%	**42.13%**	57.95%	55.08%
	MACF	78.98%	60.63%	31.60%	40.82%	57.18%	53.84%
Statistical	Sequential	70.07%	53.56%	29.24%	36.91%	57.98%	49.52%
	Random	81.00%	65.01%	32.03%	39.48%	57.77%	55.06%
	LBF	82.40%	65.34%	**33.17%**	41.78%	57.49%	56.04%
	MAEF	***83.37%***	***65.60%***	*32.74%*	*41.93%*	***58.27%***	***56.38%***
	MACF	81.59%	63.69%	31.90%	**42.07%**	57.44%	55.34%

TABLE IV
STRATEGY RESULT MATRIX - F-MEASURE AREA

results are highly heterogeneous, which Kuperus explains in depth in [14].

Figure 5 compares Answer Handling Strategies BasicDB and Statistical on a heterogeneous as well as a more homogeneous query (return all occurrences of "*Sri Lanka*"). The benefit of Statistical over BasicDB is more significant in the latter case. It can be concluded that in heterogeneous situations clustering and extrapolating probabilities merely on the name of an entity is less effective.

VI. DISCUSSION

In this paper, we presented and experiment with only the top of the iceberg of possible strategies for Targeted Feedback. A detailed list of proposals for more advanced strategies can be found in [14]. We like to mention three:

- **Context analysis** The presented strategies merely look at entities, their name, and their possible classifications. Analyzing the context, e.g. POS tags of the surrounding text, may be a useful technique in Answer Handler Strategies, extrapolating the given answer not only to entities having similar name, but also similar context.
- **Similarity** The Statistical strategy redistributes probabilities of entities having similar *name*. Similarity can be established on many other criteria. By defining a similarity function based on other features, other patterns of extrapolation can be achieved, both in Answer Handling as well as Question Proposal.
- **Extrapolation scope** The Statistical strategy redistributes probabilities of entities throughout the whole dataset. It is likely that it is not justified to extrapolate answers to such a wide scope. Limiting the scope of extrapolation may boost performance.

VII. CONCLUSIONS

Comparing *Recall* for regular NER to the $E_{100}(Recall)$ of PNER, it can be concluded that PNER shows a significantly bigger coverage of the correct answer, adding up to a difference over **29%** for the whole dataset. Comparing regular NER *Precision* and $E(Precision)$ of PNER for the whole dataset, it shows that regular NER does not weigh up to weighing

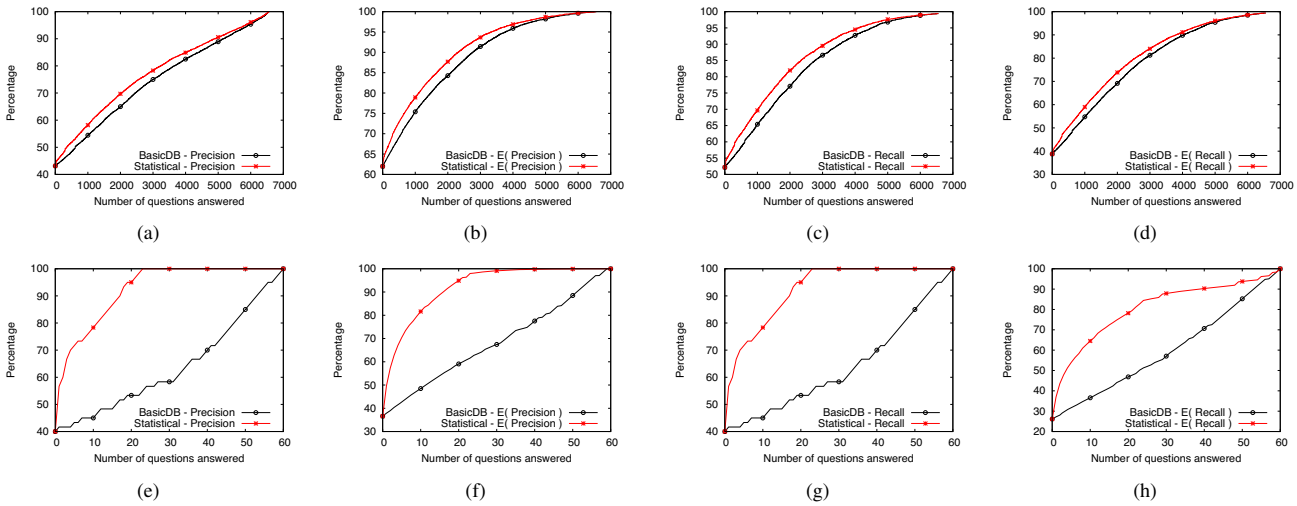

Fig. 5. BasicDB vs. Statistical using MACF; (a)-(d) Query A; (e)-(h) Query "Sri Lanka"

each correct answer by its probability. In other words, a considerable amount of entities would be lost when classified based on the highest entity class probability.

Targeted Feedback aims on solving the introduced ambiguity by posing the user targeted questions and attempting to learn from the answers provided by the user. When deploying Targeted Feedback, eventually both *Precision* and *Recall* approach 100%, proving convergence of the method. However, the time and effort of the user are costly. If the user has to inspect each reported entity before the desired data quality is reached, the user might rather choose to perform NER manually.

The implemented Targeted Feedback strategies do not show spectacular performance gain. Whether PNER is more useful than regular NER depends on the application. Clearly, there is a significant amount of ambiguity in the results. Further study is needed to find out whether the required user interaction is feasible in real forensic queries. Also secondary analysis, such as entity based document linking, must be tested for its robustness against noise.

Finally, it can be concluded that PNER in combination with Targeted Feedback shows real potential compared to regular NER. The initial PNER results cover significantly more correct answers, which would be discarded during regular NER, and using Targeted Feedback, the introduced ambiguity can be resolved and data quality in terms of both *Recall* and *Precision* eventually approach 100%.

REFERENCES

[1] D. F. Bon, "Domain specific named entity recognition for forensic intelligence," 2011, University of Amsterdam.

[2] N. Wacholder, Y. Ravin, and M. Choi, "Disambiguation of Proper Names in Text," in *Proc. of the 5th Conf. on Applied Natural Language Processing*, 1997.

[3] M. van Keulen and M. B. Habib, "Named Entity Extraction and Disambiguation: The Reinforcement Effect," in *Proc. of the 5th Int'l Workshop on Management of Uncertain Data, MUD 2011*, 2011.

[4] R. Klinger and K. Tomanek, *Classical Probabilistic Models and Conditional Random Fields*, ser. Algorithm Engineering Report. Dortmund University of Technology, 2007, no. TR07-2-013, ISSN 1864-4503.

[5] H. W. Wallach, "Conditional Random Fields: An Introduction," *Rapport technique MS-CIS-04-21, University of Pennsylvania*, vol. 50, 2004.

[6] R. Gupta and S. Sarawagi, "Creating probabilistic databases from information extraction models," in *Proc. of the VLDB Endowment*, 2006.

[7] GeoNames, "GeoNames," Available at: http://www.geonames.org/, Last visited Nov. 2011.

[8] J. Hoffart, M. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust Disambiguation of Named Entities in Text," in *Proc. of EMNLP*, 2011.

[9] Max Planck Institute for Informatics, "AIDA Web Interface," Available at: https://d5gate.ag5.mpi-sb.mpg.de/webaida/, Last visited Nov. 2011.

[10] M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum, "AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables," in *Proc. of the VLDB Endowment*, 2011.

[11] Stanford University: Natural Language Processing, "Stanford Named Entity Recognizer," Available at: http://nlp.stanford.edu/software/CRF-NER.shtml, Last visited Mar. 2012.

[12] J. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," in *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.

[13] J. Kalbfleisch, *Probability and Statistical Inference, Volume 1: Probability*. Springer, 1985.

[14] J. Kuperus, "Catching criminals by chance: a probabilistic approach to named entity recognition using targeted feedback," Master's thesis, University of Twente, 2012, http://eprints.eemcs.utwente.nl/21948.

[15] M. Reynaert, N. Oostdijk, O. De Clercq, H. Heuvel, and F. Jong, "Balancing SoNaR: IPR versus processing issues in a 500-million-word written Dutch reference corpus," in *Proc. of the 7th Conf. on International Language Resources and Evaluation (LREC)*, 2010.

[16] Nederlandse Taalunie, "STEVIN: Spraak- en Taaltechnologische Essentiële Voorzieningen in het Nederlands," Available at: http://taalunieversum.org/taal/technologie/stevin/, Last visited Oct. 2011.

[17] Centre for Language and Speech Technology, "SoNaR: STEVIN Nederlandstalig Referentiecorpus," Available at: http://lands.let.ru.nl/projects/SoNaR/, Last visited Oct. 2011.

[18] I. Schuurman, V. Hoste, and P. Monachesi, "Interacting Semantic Layers of Annotation in SoNaR, a Reference Corpus of Contemporary Written Dutch," in *Proc. of the 7th Int'l Conf. on Linguistic Resources and Evaluation (LREC)*, 2010.

[19] D. Jurafsky and J. H. Martin, *Speech And Language Processing*. Pearson Education, 2009.

[20] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern Information Retrieval*. ACM press New York, 1999.

[21] M. van Keulen and A. de Keijzer, "Qualitative Effects of Knowledge Rules and User Feedback in Probabilistic Data Integration," *The VLDB Journal*, 2009.