

Entity Models for Trigger-Reaction Documents

Mahboob Aalam Khalid* Maarten Marx Marc X. Makkes
mahboob@science.uva.nl marx@science.uva.nl mmakkes@science.uva.nl

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam

ABSTRACT

We define the notion of an entity model for a special kind of document popular on the web: an article followed by a list of reactions on that article, usually by many authors, usually inverse chronologically ordered. We call these documents trigger-reactions pairs. The entity model describes which named entities (persons, organizations, locations, products, urls) are mentioned, their type, how often and where they are mentioned, and it lists all variants referring to the same entity. These models find applications in media-analysis, trend watching, entity tracking and marketing.

The two main challenges for creating entity models are 1) detecting the entities and 2) normalizing all variants to the same correct canonical form. This task is particularly hard for user generated content on the web, of which our reactions are an example.

We use an algorithm for named entity recognition and normalization (NEN) tailor-made for trigger-reaction documents. It achieves high recall and reasonable precision by using two simple facts: 1) incomplete entities in reactions often occur complete in the trigger and 2) entities mentioned in news-articles on the web often have a Wikipedia page.

This article describes our experience in creating and using entity models on a corpus of 56.449 Dutch trigger-reaction documents, with a total of 616.715 reactions, collected from the web from November 11, 2006 to February 5, 2008. This article accompanies an earlier article from our group (currently under submission) in which the focus was on a systems-evaluation of the NEN algorithm.

Categories and Subject Descriptors

H.3.1 [Information Systems]: Information Storage and Retrieval—*Content Analysis and Indexing*

*Mahboob Aalam Khalid is supported by NWO grant 612.066.512. Thanks are due to Jur Dordrecht.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DIR-2008 April 14-15, 2008, Maastricht, the Netherlands.
Copyright 2008 by the author(s).

Keywords

Named entity normalization, Wikipedia, Opinion mining, Wikification, Language technology

1. INTRODUCTION

The task of record linkage (RL) is generally to find entries that refer to the same entity in different data sources. This task has been investigated since the 1950s, usually in the case entries are considered with their attributes (e.g., person with phone, address) [23]. This task received significance because data sources have different ways of referring to the same real-world entity due to, e.g., different naming conventions, misspelling or use of abbreviation. The task of reference normalization is to analyse and detect these different references [5, 6].

When we consider the special case of this problem for natural language texts, we first have to recognize entities in the text, and then resolve these references either to entities existing within the document or to real world entities. This, together, is the problem known as *named entity normalization* (NEN) [3, 4].

We study a type of document which is increasingly common on the web: an edited text followed by a list of usually moderated but unedited reactions to the triggering text left behind by multiple readers. We consider the pair consisting of the triggering article and the following list of reactions as one document, a trigger-reaction pair. These kind of documents occur in several contexts on the web:

weblogs: the trigger is a blog-post.

online news: the trigger is a news-story.

discussion forums: the trigger is, well, an article triggering the discussion.

online manuals: (e.g., PHP and MySQL) the trigger is a page describing a piece of the software, often a function.

In this article we only discuss documents coming from online versions of existing newspapers.

These trigger-reactions documents have a number of characteristics which make them special:

Multiple authors: The trigger and each reaction can have its own author. Authors are always known and often uniquely identifiable. This aspect may cause very diverse language use within one document.

Explicit structure: The document contains explicit meta-data. The trigger and each reaction is identifiable and both have author names, timestamps and sometimes permalinks. Authors of reactions also may attach a city and an email address.

Implicit structure: Reactions may refer to earlier reactions, which is then implicit in the text (e.g., *Bravo Henk* or *@willem*). Reactions often co-refer to named entities mentioned in the trigger or in earlier reactions.

[19] describes how the implicit discussion structure can be extracted from these documents. The reactions contain many co-references to entities mentioned in the trigger. Interestingly, the usual anaphora like pronouns are hardly used, but instead reactors use variants on the name of the entity. This is the reason that we can use named entity recognition and name-matching techniques to make those co-references explicit.

For each document we create an *entity model*. The entity model describes which named entities (persons, organizations, locations, products, urls) are mentioned, their type, how often and where they are mentioned, and it lists all variants referring to the same entity. Table 2 contains an example.

Entity models find applications in media-analysis, trend watching, entity tracking and marketing. The low response on face-to-face and telephone questionnaires makes analysts turn towards user-generated content on the web. The topicality of trigger-reaction documents makes them especially well suited for automatic analysis. Moderated sites and sites on which reactors have to login are preferred because of 1) absence of spam, and 2) reactor-models can be built reflecting the population of reactors and giving an indication of the bias in the sample [10].

Our corpus currently consists of 56.449 Dutch trigger-reaction documents (containing at least 1 reaction), with a total of 616.715 reactions, collected from the web from November 11, 2006 to February 5, 2008. The part of the corpus we use in this paper consists of documents from the online versions of four Dutch daily newspaper with a nationwide distribution: Algemeen Dagblad (AD), NRC Handelsblad, Telegraaf en Trouw. Table 1 gives some basic data about the corpus. [19] has done a qualitative analysis of this corpus.

Table 1: Description of the corpus. The mean reaction/reactors ratio is $(\sum_n(reactions(n)/reactors(n)))/n$.

	AD	NRC	Telegraaf	Trouw
Av. # of reactions per article	20.96	107.89	57.2	21.91
Av. # of reactors per article	18	62.48	55.33	20.28
Mean reaction/reactors ratio	1.1	2.23	1.02	1.07
Total # of documents	11,407	227	5,971	551

This article accompanies an earlier article from the ILPS group [12] (currently under submission) in which the focus was on a systems-evaluation of the NEN algorithm. In the current article we describe what possible uses entity models have when applied to a large set of trigger-reactions docu-

ments. Our main contribution is the list of example use-cases that we present.

The paper is organized as follows. Section 2 gives a high level description of the named entity recognition and normalization algorithm that we have developed and whose results we use in the paper. Section 3 presents a number of use-cases for entity models. We end with a section on related work followed by our conclusions and directions for future work.

2. NAMED ENTITY NORMALIZATION ALGORITHM

We now present a high level description of the algorithm from [12]. It consists of 5 separate steps. We briefly describe each step and exemplify what happens in each step using the article discussed in the next section.

Step 1. After tokenization, the named entity recognizer for Dutch made by Erik Tjong Kim Sang [21] is applied to the trigger-reactions document. This returns phrases consisting of series of tokens recognized as one named entity. Examples are *Paul de Leeuw*, *Anneke Grönloh* and *Koefnoen Spijkerman*.

Step 2. The named entity (NE) recognizer was trained on nice clean edited text and makes many mistakes in finding the correct borders of the entities. In this step, the NE output is trimmed. Examples are *Wilders!!!!!!!* and *Wilders.Ik* to *Wilders* and *mv. Gronloh* to *Gronloh*.

Step 3. Once the algorithm finds the right borders, it tries to resolve the NE in the article's own local list of normalized entities. This list contains normalized entities found earlier in the same article. The normalized form of an NE is found by using the person name matching algorithm described in [11]. It should be noted that the algorithm searches the local list in reverse order when processing the news article. This is because the last entry in the article is most likely the one being referred to.

In both this step and the next, matching of names is not strict and uses heuristics (e.g., persons have a first and last name; in Dutch words like *Mevr* are not part of a name, etc).

Step 4. If the NE is not found in the local list of normalized entities, then the NE is looked up in a list of all titles of Dutch Wikipedia articles. If the matched article is a Wikipedia redirect then we consider its target article's title instead. Then we check whether the selected article is a Wikipedia disambiguation page or not; if not then we take the article's title as normalized entity. Otherwise the Wikipedia article which has the highest number of in-links in Wikipedia is selected.¹ If the entity is not found in the local list and Wikipedia database it is taken as its own normal form, and added to the local list.

Step 5. After having created a list of normalized and recognized NE's, we go again through the reactions and find ngrams not recognized as an NE, but which partly match entities or their normalized form in the list. This step mainly recovers non-capitalized names as *anneke* and *de leeuw* (See Table 2).

¹This is inspired from Google's PageRank and also followed by [11]. In future work we will improve on this by comparing the language used in the document to the language used in the various Wikipedia pages, and choosing the most similar page.

Table 2: The entity model of an article about Paul de Leeuw and Anneke Grönloh, <http://www.telegraaf.nl/prive/article74586221.ece>. Position information is omitted.

Normalized name	Type	# in trigger	# in reactions	Variants
Paul de Leeuw	PER	6	63	Paul de L, P.de L, P de Leeuw, Paul de leeuw, Paul, Leeuw, paul de leeuw, paul, de leeuw, leeuw, De Leeuw
Anneke Grönloh	PER	9	54	Anneke Gr, Anneke Gronloh, Mevrouw G, Anneke, AG, Gronloh, Gr ?nloh, anneke, gronloh Anneke Kreunlo, Anneke Grohnlloh, Anneke Kreuntzo
Nederland	PER	0	4	Nederland, Nederlandse
Jack Spijkerman	PER	0	3	Spijkerman, spijkerman
Koefnoen	MISC	0	2	Koefnoen

3. ENTITY MODELS AND THEIR USE

This section first presents entity models of one document. After that we create an entity model of one person, Geert Wilders, a controversial Dutch politician and a language model. We finish the section by looking at trends: for Geert Wilders.

3.1 Examples of entity models

Table 2 contains the entity model of an article in the Telegraaf² of November 6, 2007 having 290 reactions. Its head and opening line were

Anneke Gronloh verwittigt advocaat over persiflage Paul de Leeuw

AMSTERDAM - Anneke Grönloh heeft haar advocaat op de hoogte gesteld van de persiflage op haar die Paul de Leeuw tijdens zijn Symphonica in Rosso-concerten laat zien. "Hij is hiermee duidelijk in overtreding"

(Indeed, in the title of the original article **Gronloh** is wrongly written without the ö.)

The entity model shows several things: The reactors have clearly picked out the two main characters in the story (Paul de Leeuw and Anneke Grönloh with 63 and 54 mentionings, respectively). The number of variants is high and names are used very creatively. It is clear that recognizing these names as variants of the normalized name is a difficult task. Nevertheless, the NEN algorithm from [12] received an accuracy of 94% on this article. The accuracy of normalization was measured using the following metric:

$$accuracy = \frac{|TP|}{|TP| + |FP_{norm}|}. \quad (1)$$

Where TP is the set of true named entities (that are recognized correctly) which are normalized to right Wikipedia articles, and FP_{norm} is the set of true named entities which are normalized incorrectly.

Figure 1 partly shows the entity models of two more articles: only the top 10 most mentioned entities are listed, and only the counts in the trigger and in the reactions are given. Two things are noteworthy. First, among the top 10 entities returned by the system all have a Wikipedia page. Second, in both articles only one of the top 10 is not a named entity (Politieke Partij and Gebruiker). The figures also show that the reactions contain the relevant entities frequently, e.g., "Pim Fortuyn" has similarity with "Geert Wilders",

²<http://www.telegraaf.nl/prive/article74586221.ece>

and "Windows Vista", "Macintosh" and "Ubuntu" are in the contest of presenting elegant graphical user interface.

These scores seem rather robust. We have evaluated on 5 Dutch and 5 BBC articles and of the total of 100 top-10-entities, 99 had a Wikipedia page and 2 were not named entities. Figure 2 shows a similar behaviour: in the top 100 recognized entities in a corpus about Geert Wilders (described in section 3.2), there were 8 mistakes (not named entities) and only 9 out of 100 did not have a Wikipedia title.

%

Direct use of entity models..

The high accuracy of the NEN algorithm suggests two simple but useful applications. (1) Accurate Wikification of the named entities in the document becomes possible. Wikification denotes the process of automatically assigning hyperlinks to Wikipedia articles to phrases occurring in texts [16]. (2) The NE counts in the reactions give a good indication of the main characters in the triggering story. As such we can use this information to improve early precision of information retrieval on news-articles by expanding the documents with multiple copies of the named entities with a high reaction-count.

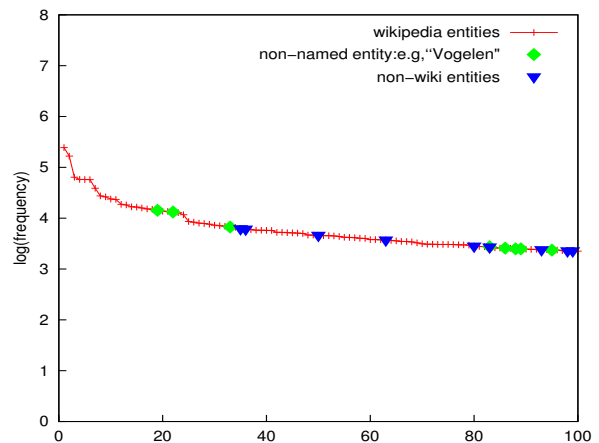
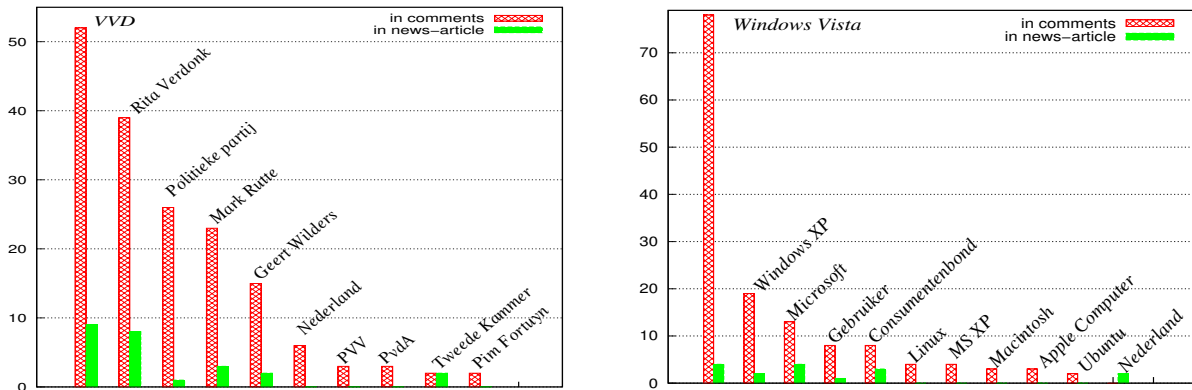


Figure 2: Frequencies (in log-scale) of the top 100 most mentioned recognized and normalized entities in the Geert Wilders corpus. Entities without a Wikipedia page and mistakes are indicated.

Figure 1: Entity models of two AD articles (<http://www.ad.nl/binnenland/article1660760.ece>, <http://www.ad.nl/economie/article1734578.ece>) showing the the counts in both the trigger and the reactions of the top 10 and 11 most mentioned entities. All entities mentioned in the article are displayed.



3.2 Personal language models

Once we have identified and normalized named entities in a large corpus we can do all kinds of interesting analysis tasks. A task of particular interest in media-analysis is to create models of persons or products based on the language used in reactions about them [10].

As an example we created an entity model of Geert Wilders as follows: we selected all news-articles which mentioned Geert Wilders, and collected all reactions belonging to those; of this total corpus we created an entity model just as we did before for one article. The entity models of news articles about Geert Wilders (Dutch politician) and their reactions were generated by taking news articles of April 2, 2007 – February 2, 2008 collected from the news websites. We have made only the entity part of the data available for research. You can download the data from <http://staff.science.uva.nl/~mahboob/dir2008/>.

The top-40 most mentioned named entities in all these reactions are given in Figure 3. The entity model closely reflects the profession of Geert Wilders (politician), and his main political issue (integration and islam related). It is noteworthy to see that not all entities are resolved to Wikipedia pages and that two variants of Geert Wilders were not normalized correctly (“Geert WildersA” and “wildERS”). We have found there is a entity “Socialitsche Partij Anders” occurs in top 15, that is a strange because this entity is talking about Belgium political party and the corpus contains reactions from Dutch commentors. This is normalized form of the named entity “SP” whose frequency is 1726 in the corpus, and according to Wikipedia knowledge³ the entity “SP” may refer to “Socialitsche Partij Anders” or “Socialistische Partij (Nederland)”, but according to our disambiguation method (see Section 2) “Socialitsche Partij Anders” was selected because it has high number of inlinks. This error could be tackled when we consider the context of a named entity, and this is our future work.

3.3 Trend watching

As a last application we mention following issues related to a named entity through time. In the entity models given

³We have used the Dutch Wikipedia collection as of 6/11/2006

above we have added all reactions about one entity given over an extended period of time. In Figure 4 we show how the frequencies of important named entities in the model of Geert Wilders change over time. The counts are based on the same dataset mentioned earlier. The graph clearly shows the ability of Geert Wilders to remain constantly in the news and being discussed. The peak detection and explanation techniques used in the MoodViews system [17] can be applied here as well.

4. RELATED WORK

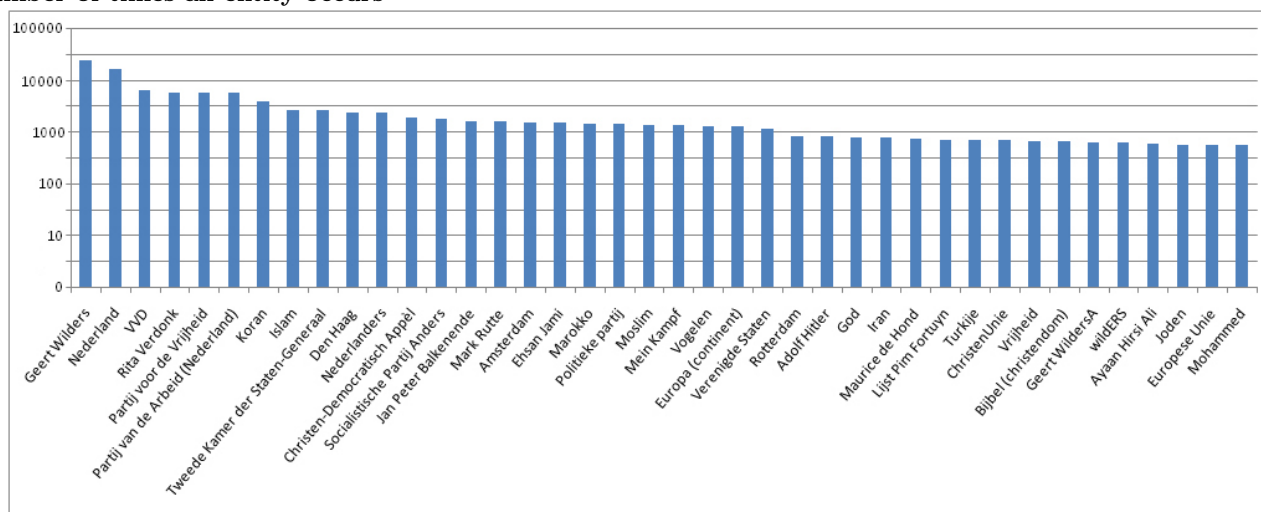
Name matching and disambiguation has been recognized as an important problem in various domains. Borgman and Siegfried [2] present an overview of motivations, applications, common problems and techniques for name matching in the art domain; see [18] for recent experiments with classification for name matching. A dual problem, personal name disambiguation has also attracted a lot of attention, and a number of unsupervised methods were proposed [15, 20].

A similar problem has been known in the database community for over five decades as the *record linkage* or the *record matching* problem [7, 23]. However, there the task is more general: matching arbitrary types of records, not just person names or other types of named entities. Another type of research focuses on identification, disambiguation and matching of text objects other than named entities, specifically, temporal expressions [1].

Related problems occur in a different task: discovering links in text, or *wikifying*. Like NEN, this task involves identifying and disambiguating references to entities, and has also attracted attention of the research community [9, 16].

Research on named entity extraction and normalization has been carried out in both restricted and open domains. For example, for the case of scientific articles on genomics, where gene and protein names can be both synonymous and ambiguous, Cohen [3] normalizes entities using dictionaries automatically extracted from gene databases. For the news domain, Magdy et al. [14] address cross-document Arabic person name normalization using a machine learning approach, a dictionary of person names and frequency information for names in a collection. Cucerzan [4] considers the

Figure 3: Named entity model of Geert Wilders. X-axis denotes the top 100 most mentioned named entities in reactions attached to articles about Geert Wilders taken from April 2, 2007 to Feb 2, 2008 (2,671 articles and 48,898 reactions). Names used are the titles of the Wikipedia pages of these entities (except for VVD) or (in case there is no Wikipedia page) entities as they occur in the text. Y-axis (in logscale) denotes the number of times an entity occurs



entity normalization task for news and encyclopedia articles; they use information extracted from Wikipedia combined with machine learning for context-aware name disambiguation. [4] also presents an extensive literature overview on the problem.

Recent research has also examined the impact of normalizing entities in text on specific information access tasks. Zhou et al. [24] show that appropriate use of domain-specific knowledge base (i.e., synonyms, hypernyms, etc.) yields significant improvement in passage retrieval in the biomedical domain. Similarly, Khalid et al. [11] demonstrate that NEN based on Wikipedia helps text retrieval in the context of Question Answering in the news domain.

Finally, in recent years, there has been a steady increase in the development or adaptation of language technology for user generated content. Most of the attention has gone to blogs (see [17] for a recent survey on text analytics for blogs). Online discussion fora are more closely related to the data with which we work; recent research includes work on finding authoritative answers in forum threads [8, 13], as well as attempts to assess the quality of forum posts [22]. To the best of our knowledge, discussion threads as triggered by news stories of the kind considered here have not been studied before.

5. CONCLUSION AND FUTURE WORK

We can conclude that named entity normalization using Wikipedia works well for Dutch trigger-reaction documents coming from online versions of daily newspapers. The algorithm from [12] achieves an average accuracy of 89%. Here we have shown that normalizing to Wikipedia URLs works for the most frequent entities: almost all have a Wikipedia page. We have shown that careful NE normalization is useful for document expansion, wikification, and creating language models of named entities, both static and diachronic.

There remain several interesting directions for follow-ups on this work. We believe that the precision of the NEN

algorithm can be improved by 1) NE disambiguation using a comparison of the language model of the document with that of the candidate Wikipedia page (e.g., *Apple* in a reaction on an article about *Windows Vista* should not be normalized to *Apple records*); 2) improved clustering of Wikipedia pages in NON-NE and NE pages, and improved typing of the latter (in the categories person, location, and organization); 3) using statistical information about NE's (in the current document, in the current period) for normalization of variants which are syntactically far removed from the normalized form.

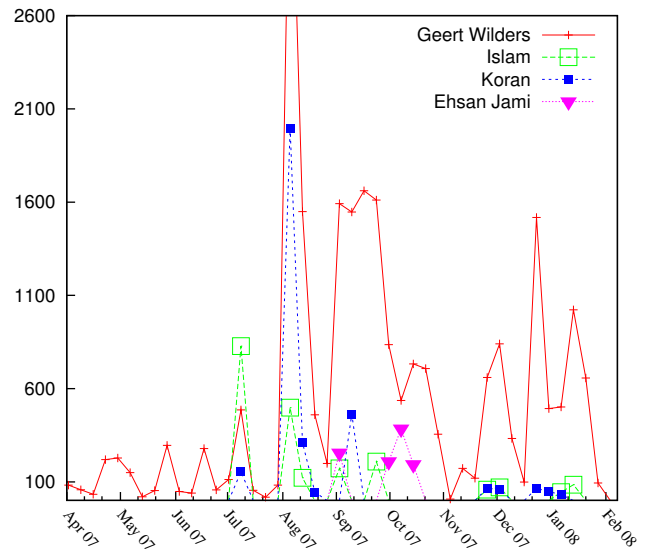
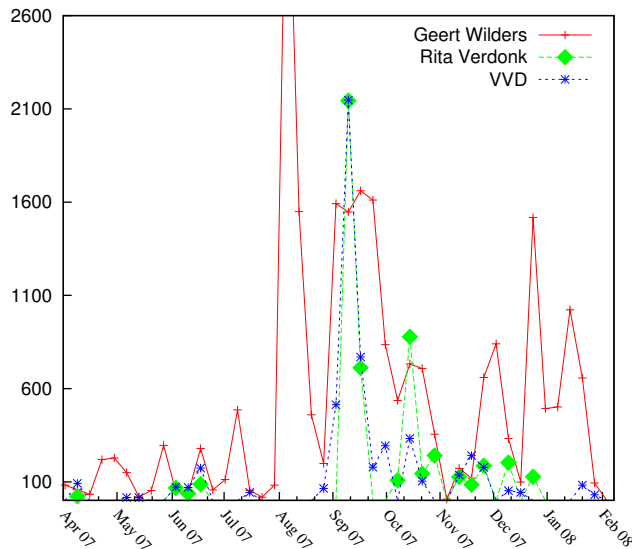
The NEN algorithm has been evaluated on relatively clean user generated content coming from moderated websites connected to newspapers with a long standing reputation. The reactions on purely online news-sites as *nu.nl* or group-logs as *geenstijl.nl* have more challenging use of language, punctuation and style. Adjusting the NEN algorithm to those cases is a challenge. Discussion fora with complicated quotation and linking possibilities provide further challenges.

Future work should also address computational issues: creating language models for entities and showing trends should be possible without delay and in query-time. Pre-computing and caching results for popular queries and materializing computed results are probably needed.

6. REFERENCES

- [1] D. Ahn, J. van Rantwijk, and M. de Rijke. A cascaded machine learning approach to interpreting temporal expressions. In *Human Language Technologies 2007: Proc. ACL 2007*, pages 420–427, 2007.
- [2] C. L. Borgman and S. L. Siegfried. Getty's synonym and its cousins: A survey of applications of personal name-matching algorithms. *Journal of the American Society for Information Science*, 43(7):459–476, 1992.
- [3] A. M. Cohen. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*, pages 17–24, 2005.

Figure 4: Wilders through time April 2, 2007 to Feb 2, 2008



- [4] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL '07*, pages 708–716, 2007.
- [5] A. Doan and A. Y. Halevy. Semantic-integration research in the database community. *AI Mag.*, 26(1):83–94, 2005.
- [6] X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *Proc. SIGMOD '05*, pages 85–96, 2005.
- [7] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.
- [8] D. Feng, E. Shaw, J. Kim, and E. Hovy. Learning to detect conversation focus of threaded discussions. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 2006.
- [9] S. Fissaha Adafre and M. de Rijke. Discovering missing links in Wikipedia. In *Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)*, 2005.
- [10] R. Franz. Personal communication. Trendlight BV, 2007.
- [11] M. Khalid, V. Jijkoun, and M. de Rijke. The impact of named entity normalization on information retrieval for question answering. In *Proc. ECIR 2008*, 2008.
- [12] M. Khalid, V. Jijkoun, M. Marx, and M. de Rijke. Named entity normalization in user generated content. Under submission, 2008.
- [13] J. Kim, G. Chern, D. Feng, E. Shaw, and E. Hovy. Mining and assessing discussions on the web through speech act analysis. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference*, 2006.
- [14] W. Magdy, K. Darwish, O. Emam, and H. Hassan. Arabic cross-document person name normalization. In *CASL Workshop '07*, pages 25–32, 2007.
- [15] G. S. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*, pages 33–40, 2003.
- [16] R. Mihalcea and A. Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 233–242, 2007.
- [17] G. Mishne. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, Amsterdam, 2007.
- [18] C. Phua, V. Lee, and K. Smith. The personal name problem and a recommended data mining solution. In *Encyclopedia of Data Warehousing and Mining (2nd Edition)*. 2006.
- [19] A. Schuth, M. Marx, and M. de Rijke. Extracting the discussion structure in comments on news-articles. In *9th ACM International Workshop on Web Information and Data Management (WIDM 2007)*, November 2007.
- [20] Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 2007 Joint Conference on Digital Libraries*, pages 342–351, 2007.
- [21] E. F. Tjong Kim Sang. Memory-based named entity recognition. In *Proceedings of CoNLL-2002, Taipei, Taiwan*, pages 203–206, 2002.
- [22] M. Weimer, I. Gurevych, and M. Mehlhauser. Automatically assessing the post quality in online discussions on software. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 125–128, 2007.
- [23] W. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC., 1999.
- [24] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *SIGIR '07*, pages 655–662, 2007.