Named Entity Tagging a Very Large Unbalanced Corpus: Training and Evaluating NE Classifiers

Joachim Bingel, Thomas Haider
Institute for the German Language (IDS)
R5 6–13, 68161 Mannheim, Germany
{bingel|haider}@ids-mannheim.de

Abstract

We describe a systematic and application-oriented approach to training and evaluating named entity recognition and classification (NERC) systems, the purpose of which is to identify an optimal system and to train an optimal model for named entity tagging DEREKO, a very large general-purpose corpus of contemporary German (Kupietz et al., 2010). DEREKO's strong dispersion wrt. genre, register and time forces us to base our decision for a specific NERC system on an evaluation performed on a representative sample of DEREKO instead of performance figures that have been reported for the individual NERC systems when evaluated on more uniform and less diverse data. We create and manually annotate such a representative sample as evaluation data for three different NERC systems, for each of which various models are learnt on multiple training data. The proposed sampling method can be viewed as a generally applicable method for sampling evaluation data from an unbalanced target corpus for any sort of natural language processing.

Keywords: Named entity recognition, Evaluation, Very large corpora

1. Introduction

The German Reference Corpus DEREKO, a very large general-purpose corpus of contemporary German texts, serves as a "primordial sample from which virtual corpora can be drawn for the specific purposes of individual studies" (Kupietz et al., 2010). The corpus is accessible via the COSMAS II corpus analysis system, which currently has more than 30,000 registered users. DEREKO's importance in German linguistics makes it desirable to provide researchers with access to named entity annotations in the corpus, which is the outset for this study. While the texts in DEREKO are (partly) tagged for morphological information, the absence of named entity annotations calls for a processing of the corpus with a NE tagger.

DEREKO's constant and rapid growth[1] as well as its strong dispersion wrt. genre, register and time force us to base our decision for a specific NERC system on an evaluation performed on a representative sample of DEREKO instead of performance figures that have been reported for the individual NERC systems when evaluated on more uniform and less diverse data. To tackle this issue, we create and manually annotate such a representative sample as evaluation data for three different NERC systems, for each of which various models are learnt on multiple training data.

Note that this study is not so much an exhaustive evaluation of a broad array of NERC systems, nor is its central purpose to judge which of the systems performs best in a general setting. Instead, we want to describe the process of selecting an optimal system for the purpose of tagging a very large multi-domain corpus and to outline all necessary steps that are involved in this process.

[1]The corpus grew from five billion to six billion between 2012 and 2013. For 2014, a growth by factor four is estimated (Kupietz and Lüngen, 2014).

2. Named entities

2.1. NE tagging of DEREKO

The identification and classification of names and named entities plays a central role in many different linguistic and text-processing applications such as question answering (Mollá et al., 2006), machine translation (Babych and Hartley, 2003), text classification (Gui et al., 2012) or even in the life sciences, for instance bioinformatics (Humphreys et al., 2000). Beyond this, quantitative linguistic research such as discourse analysis may profit from named entity information (van Vliet, 2008) or may be interested in exploring the syntactic behaviour of named entities or their frequency in specific genres and domains. Hence, enriching DEREKO, the largest systematic collection of German texts, is an obvious benefit.

Most NERC systems expect a single text genre and domain as training and processing input. In the case of older systems using handcrafted rules, genre is usually limited to newswire text (Poibeau and Kosseim, 2001), typically in the domain of politics, business or sports. In fact, the impact of genre and domain has not been thoroughly studied in the NERC literature (Nadeau and Sekine, 2007). The dispersion over text genre[2] is a crucial bottleneck for our enterprise of NE tagging DEREKO, for as Poibeau and Kosseim (2001) are able to show, performance may drop drastically when a system trained on newswire text is ported to a different domain.

Later systems, which are usually based on machine learning techniques, view named entity recognition similar to more straightforward NLP tasks such as PoS tagging. They employ n-grams, chunking information and orthographic information (e.g. capitalisation). As Tjong Kim Sang and de Meulder (2003), note, none of the features works well for named entity recognition when used in isolation. Zhang and Tong (2004) study the impact of certain features on

[2]Although the largest portion of the texts in DEREKO is actually acquired from newspapers, the corpus also contains fiction, web texts and political speeches.

Label	Meaning
PER	Person
ORG	Organisation
LOC	Geographic location
GPE	Geopolitical entity[3]
OTH	Other

Table 1: Named entity classes used in this study.

several NERC algorithms and discover conditional random fields (CRF) to be particularly useful because of their ability to detect semantic similarities between NE candidates. For instance, *Barack Obama* is often accompanied by certain lexemes which are indicative of a U.S. president. It is therefore assumed that the classification algorithm will be able to infer that the phrase *George W. Bush* also denotes a person because the it is capitalised and occurs in a similar context as *Barack Obama*. The Stanford NERC system, which we use in this study, makes use of this method.

2.2. NE class set for this study

As for several other linguistic categories, there is no fixed and final set of classes for named entities, although PERSON, ORGANISATION or LOCATION tend to be present in most class sets. Beyond these, the class set used for the CoNLL-2003 NER shared task (Tjong Kim Sang and de Meulder, 2003) contains a MISCELLANEOUS class, and the classes DATE, TIME, MONEY and PERCENT have been used for MUC-6 and MUC-7. Other common classes include GEOPOLITICAL ENTITY and QUANTITY. Nadeau and Sekine (2007) provide an overview of the class sets used for various shared tasks and demonstrate their heterogeneity.

For the sake of comparability of the various resources used in this study (NE taggers as well as gazetteers and corpora), we use a small set comprising the classes listed in table 1. The same set is used in the Tüba-D/Z treebank (Telljohann et al., 2004), our main training resource.

3. Evaluation data

3.1. DEREKO sample creation

The DEREKO-2013-II release (which we used for sample creation) comprises more than 6 billion words that are unevenly distributed across genre (although newspaper texts clearly make for the strongest stratum) and time (although relatively few texts reach back to the mid-20th century). As outlined above, we want to draw a sample that is as representative of DEREKO as possible. However, we want to keep the annotation effort low, i.e. the sample should not be larger than necessary. We are thus faced with the problem of finding the minimally required sample size n. Independently from population size, n is a function of an arbitrary margin of error e and an arbitrary confidence level α by which a statistics on the sample does not deviate from the same statistics on the population by more than e (Bortz and Döring, 2006). We postulate $e = 1.5\%$ and $\alpha = 99\%$ and determine n using a z-transformation:[4]

$$n = \frac{z^2_{\alpha/2}\sigma^2}{e^2\sigma^2} = \frac{2,58^2}{0,015^2} \approx 17,073 \quad (1)$$

Consequently, the sample needs to comprise at least 17,073 decisions to be made for the NERC systems, i.e. 17,073 tokens.

In order to reflect the distribution of words across genres in DEREKO, we want each of the 439 subcorpora[5] to have a more or less equal share in the sample. We determine the number of words to extract from every subcorpus c by computing

$$w(s_c) = \left\lceil \frac{w(c) \cdot n}{w(D)} \right\rceil \quad (2)$$

where n is the previously computed 17,073, s_c is the subcorpus sample, D is DEREKO and w is the word count function for a text collection. Naturally, we extract full randomly chosen sentences from the subcorpora instead of single words. We do this iteratively for every subcorpus until its required word count is reached. The sample ultimately consists of 18,079 tokens (1,092 sentences) representing the 439 subcorpora.

3.2. Annotation

We manually annotate the evaluation sample using GATE (Cunningham et al., 2013) according to certain guidelines. These are established with respect to the algorithms and data formats used by the examined NERC systems, with the ultimate goal to ensure a maximal comparability between manual annotation and system outputs.

Organisations with place names

A very common conflict arises in the case of organisations that carry a place in their name, e.g. sports teams. An example from the evaluation data is *TV Bad Ems*, denoting a sports club from the town of Bad Ems. Since a nested analysis like `<ORG>TV <GPE>Bad Ems</GPE></ORG>` is not supported by all systems, we define the guideline that we always annotate the maximal span or highest-level entity, in this case the organisation. The place name *Bad Ems* is consequently not annotated.

Organisations denoted by common nouns

A similar case is that of organisations that are denoted with German common nouns. In the evaluation sample, examples of this are *Landesbezirk Villingen* (state district Villingen) or *Universitätsklinikum Ulm* (university hospital Ulm). We examine these cases individually and, depending on the respective context, decide which entity (the organisation or the subsequent place name) is more prominent. Thus, we annotate the mentioned examples `Landesbezirk <GPE>Villingen`

[3] A GPE differs from a LOC through its political dimension. Nadeau and Sekine (2007) ascribe some sort of government to a GPE.

[4] For a more detailed explanation of the formula as well as the z-transformation, see (Bortz and Döring, 2006).

[5] The DEREKO-2013-II release is organised in 439 files, which we perceive as subcorpora. A file usually corresponds to a volume of a certain newspaper or a collection like *bibliographies published in the* 20^{th} *century.*

(Baden)</GPE> and <ORG>Universitätsklinikum Ulm</ORG>, respectively.

Place names used as adjectives
Entities (organisations or persons) that are modified with place names in an adjectival manner in German are evident in the following examples: *Mannheimer Stadtverwaltung* (Mannheim city council), *schleswig-holsteinischer Hotel- und Gaststättenverband* (Hotel and Restaurant Association of Schleswig-Holstein) or *die Partenkirchnerin Maria Riesch* (Maria Riesch, a skiing professional from Partenkirchen). In these cases, we do not annotate place names in the evaluation data in order to ensure consistency with the training data, in which this problem is treated likewise.

Coordination of composed proper names
In the evaluation sample, we observe in *Lani- und Gelingenbach* one example of 'N- and N+N' coordination of proper names (which are originally noun compounds). This case is problematic because it is unclear whether and how the first coordinated name (Lani-[bach]) can be annotated. As we consider this case highly difficult for any NERC system to solve and also to be extremely exceptional[6], we remove the containing sentence from the data.

3.3. Distribution of named entities in the evaluation data

The evaluation sample comprises a total of 1,454 annotated named entities that distribute across the classes as listed in table 2.

Class	PER	ORG	LOC	GPE	OTH	Total
Instances	652	401	130	230	41	1,454

Table 2: Distribution of named entities across classes

4. Experiments

4.1. Systems and training data

We train different models for the Stanford NERC (coreNLP) system v3.2.0 (Finkel et al., 2005), the JULIE Named Entity Tagger [JNET] (Hahn et al., 2008), as well as Apache OpenNLP[7]. The former two systems are based on Conditional Random Fields, while the latter is a maximum entropy classifier. As training data, we use Tüba-D/Z, a newspaper corpus comprising 75k sentences (1.3M tokens) that are manually annotated for named entities, parts-of-speech, coreference relations and various syntactic information. In addition to Tüba-D/Z, we evaluate two models developed for the Stanford system by Faruqui and Padó (2010), trained on the Huge German Corpus [HGC] (Schmid, 1994; Schiller et al., 1999) and on the deWaC corpus (Baroni et al., 2009), respectively. In another experiment on the impact of domain differences (see section 5.2.3.), we use the Europarl corpus (Koehn, 2005) as testing data.
We first gauge the performance of a system by evaluating it against the DEREKO sample, and in a second series of experiments unify the locality classes LOC and GPE in order to determine the difficulty of the systems in distinguishing the two.

4.2. Gazetteers

In order to back up the systems with pre-compiled lists of named entities, we extract names of persons, organisations and places from DBpedia (Bizer et al., 2009), and further place names from the GeoNames database (GeoNames, nd), distinguishing between GPE and LOC according to the feature codes used in GeoNames.[8] We extract about 75k proper names for every class from DBpedia and 200k German place names from GeoNames.

5. Evaluation of systems

5.1. Evaluation metrics

A common metric for the evaluation of NERC systems is the MUC score, which distinguishes between so-called TEXT predictions and TYPE predictions (Nadeau and Sekine, 2007), where the former relates to the correct identification of a text span as a named entity (recognition task) and the latter to the correct assignment of a class label to identified named entities (classification task). Standard evaluation metrics such as recall, precision and f-measure can then be applied individually to TEXT and TYPE predictions. This method has a clear advantage compared to an exact matching of TEXT and TYPE in that a system that correctly identifies many named entities but assigns wrong class labels is valued higher than a system that does not find these named entities at all.
We test differences in evaluation results for statistical significance using approximate randomisation tests (Noreen, 1989; Yeh, 2000). We compute the test statistics using the sigf package (Padó, 2006).

5.2. Experimental results

Table 3 shows the TEXT- and TYPE-based F_1-scores for the systems trained on the Tüba-D/Z. We observe an equal performance of the Stanford and JNET systems wrt. the TEXT statistics and for the TYPE statistics a slight but not highly significant superiority of the Stanford system ($p = 0.064$). The OpenNLP system lags behind in both statistics.

System	F_{TYPE}	F_{TEXT}
Stanford	.689	.790
JNET	.673	.790
OpenNLP	.446	.579

Table 3: F_1-scores for TYPE and TEXT for NERC systems trained on Tüba-D/Z

[6] We do not expect a significant number of other occurrences of such cases in the corpus, as would be *Switzer- and Finland* or *Liver- and Blackpool*.

[7] http://opennlp.apache.org

[8] We utilise the top- level feature codes, e.g. *A* for country, state, region (GPE) or *H* for a stream or lake (LOC).

Class	Precision	Recall	F_1
PER	.829	.738	.780
ORG	.630	.355	.454
GPE	.729	.715	.722
LOC	.477	.287	.358
OTH	.143	.024	.041
Total	.728	.563	.689

Table 4: Stanford Tagger performance by NE class

Table 4 displays the Stanford system's performance as recall, precision and F_1 by NE class. We observe particular difficulties in identifying ORGs and LOCs, which may be partly due to an interference of the classes as outlined in the case of organisations with places in their names (cf. section 5.3. on error analysis). Additionally, only one of the 41 NEs labelled OTH is recognised. The results reflect the general observation (which also holds for the other systems and settings) that precision is considerably higher than recall.

5.2.1. Simplified classification task

We train and test further models for 4 instead of 5 classes, unifying LOCs and GPEs in order to avoid the systematic difficulties in distinguishing these classes and also in order to allow for a comparison with the HGC and deWaC models, which only comprise the 4 remaining classes. Table 5.2.1. shows the results for a reduced class set. Again, we observe a superiority of the Stanford system and in particular of the Tüba-D/Z model towards the other models. The evaluation statistics is significantly higher for the winning system ($p < 0.01$).

System	F_{TYPE}	F_{TEXT}
$\text{Stanford}_{\text{Tüba-D/Z}}$	.721*	.800*
$\text{JNET}_{\text{Tüba-D/Z}}$	.691	.783
$\text{OpenNLP}_{\text{Tüba-D/Z}}$	.446	.579
$\text{Stanford}_{\text{deWaC}}$	.612	.761
$\text{Stanford}_{\text{HGC}}$	.624	.756

Table 5: F_1-scores for 4 NE classes (PER, ORG, LOC+GPE and OTH). Subscripts indicate training data.
*=Statistically significant with $p < 0.01$.

5.2.2. Gazetteers

We further examine the influence of gazetteers on the Stanford system. Table 5.2.2. shows an improved performance using DBpedia and another small increase using GeoNames, although none of the two leads to a statistically significant improvement ($p \approx 0.45$, respectively).

System	F_{TYPE}	F_{TEXT}
Stanford	.689	.790
+DBpedia	.693	.792
+DBpedia+GeoNames	.693	.800

Table 6: F_1-scores using gazetteers, 5 NE classes

5.2.3. Robustness across domains

Finally, we are interested in the potential of the systems given equal and very different training and testing domains. This experiment is intended to shed some light on the question how dependent the individual systems are on domain consistency in training and testing. Table 7 shows the TYPE- and TEXT-based F_1-scores for the Stanford and JNET systems trained and evaluated on a 90/10 split of the Tüba-D/Z. We observe a clear and significant improvement ($p < 0.01$) compared to the DEREKO statistics and a significant superiority ($p < 0.01$) of JNET towards Stanford. For very different training and testing domains, we evaluated the Tüba-D/Z-trained models against the NE-annotated German section of the Europarl corpus (Koehn, 2005; Faruqui and Padó, 2010), yielding the results presented in table 8. Here, we observe that Stanford is clearly superior to JNET. These results lead us to say that the Stanford systems seems to be less sensitive to training/testing domain.

System	F_{TYPE}	F_{TEXT}
Stanford	.818	.888
JNET	.844*	.976*

Table 7: Performance evaluation with equal train/test domain ($\text{Tüba-D/Z}_{90\%}$ and $\text{Tüba-D/Z}_{10\%}$).
*=Statistically significant with $p < 0.01$.

System	F_{TYPE}	F_{TEXT}
Stanford	.458*	.544*
JNET	.300	.515

Table 8: Performance evaluation with different train/test domains (Tüba-D/Z and Europarl).
*=Statistically significant with $p < 0.01$.

Class	PER	ORG	LOC	GPE	OTH	¬NE
PER	**535**	7	-	3	-	107
ORG	16	**141**	10	13	1	220
LOC	11	6	**33**	15	-	65
GPE	10	13	12	**157**	-	38
OTH	2	2	1	1	**1**	34
¬NE	21	19	2	8	5	**15041**

Table 9: Confusion over classes (Stanford system trained on Tüba-D/Z, without gazetteers). Rows indicate gold annotation, columns indicate system output. ¬NE stands for tokens that are not (parts of) named entities.

5.3. Error analysis

Table 9 displays the confusion over NE classes for the basic Stanford system (trained on Tüba-D/Z, no gazetteers). We observe general difficulties for the system in *detecting* named entities, with gold NEs not identified as such by the system regardless of the correct class. In the cases of ORG, LOC and OTH, tokens are more often left undetected as NEs

than they are classified as the gold class. In particular, many of the ORG tokens are not detected as named entities when the gold annotation comprises a token span that denotes a sports team with a town name in it (which is annotated ORG as a whole according to the guidelines in section 3.2.). In these cases, the system often identified the town as LOC or GPE and disregarded the team acronym. The confusion of LOC for the conceptually very similar class GPE is relatively high compared to other class pairs, although the inverse confusion (GPE as LOC) is not as significant. In a closer analysis of the results for the experiments on different training and testing domain (cf. table 8), we noted that the Europarl named entity annotations denoted *Das Parlament* ('the parliament', a frequent phrase in the corpus) as an ORG named entity, which was not tagged by any of the systems.

6. Future Work

While this work tried to minimise annotation effort, more resources might have enabled us to compile a larger data set that would be sufficient in size to also be used for training (possibly cross-validation), employing the same sampling technique in order to maintain representativeness of DEREKO. A possible benefit of this is demonstrated by the results reported in table 7, where equal training and testing domains lead to very good results. Addressing the strong dispersion of domain and genre in DEREKO, domain adaptation techniques (e.g. LSA-based methods as suggested by Guo et al. (2009)) may prove helpful. Another alternative to overcome this problem might be to use different models for different domains and text genres.

7. Conclusion

This study presented a systematic evaluation of NERC systems with the purpose of identifying an optimal system and model for tagging DEREKO. The experimental results generally suggest using the Stanford system for the annotation of DEREKO, although the difference between Stanford and JNET is not statistically significant. However, the Stanford system appears to be more robust towards genre/domain variance, which is clearly beneficial when tagging unbalanced corpora. We further observe a very slight impact of gazetteers on the tagging performance. The very good results for the in-domain evaluation in 5.2.3. hint at a certain potential of a model that is trained on more representative training data than Tüba-D/Z, however the inherent genre variance as well as the increased annotation effort pose high obstacles for training.

The central contribution of this study is its description of our process of identifying the optimal NERC system for tagging a very large unbalanced corpus. This process involved the systematic creation and manual annotation of a representative sample of our target corpus. Overall, we view this as a generally applicable method for sampling evaluation data from an unbalanced target corpus for any sort of natural language processing.

8. References

Babych, B. and Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, pages 1–8. Association for Computational Linguistics.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia-A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165.

Bortz, J. and Döring, N. (2006). *Forschungsmethoden und Evaluation: Für Human-und Sozialwissenschaftler*. Springer.

Cunningham, H., Tablan, V., Roberts, A., and Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS computational biology*, 9(2):e1002854.

Faruqui, M. and Padó, S. (2010). Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of KONVENS*, pages 129–133.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

GeoNames. (n.d.). Geonames. http://geonames.org/. Retr. June 17, 2013.

Gui, Y., Gao, Z., Li, R., and Yang, X. (2012). Hierarchical text classification for news articles based-on named entities. In *Advanced Data Mining and Applications*, pages 318–329. Springer.

Guo, H., Zhu, H., Guo, Z., Zhang, X., Wu, X., and Su, Z. (2009). Domain adaptation with latent semantic association for named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 281–289, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hahn, U., Buyko, E., Landefeld, R., Mühlhausen, M., Poprat, M., Tomanek, K., and Wermter, J. (2008). An Overview of JCoRe, the JULIE Lab UIMA Component Repository. In Hahn, U., Götz, T., Brown, E. W., Cunningham, H., and Nyberg, E. H., editors, *Proceedings of the LREC '08 Workshop "Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP". Marrakech, Morocco, 31 May 2008*, pages 1–7.

Humphreys, K., Demetriou, G., and Gaizauskas, R. (2000). Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In *Pac Symp Biocomput*, volume 5, pages 505–516.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages

79–86.

Kupietz, M. and Lüngen, H. (2014). Recent developments in DeReKo. In *this volume*.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Mollá, D., Van Zaanen, M., and Smith, D. (2006). Named entity recognition for question answering. *Proceedings of ALTW*, pages 51–58.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Noreen, E. W. (1989). Computer Intensive Methods for Testing Hypotheses. An Introduction. *John Wiley & Sons*, 2(5):33.

Padó, S. (2006). User's guide to sigf: Significance testing by approximate randomisation.

Poibeau, T. and Kosseim, L. (2001). Proper Name Extraction from Non-Journalistic Texts. In *Computational Linguistics in the Netherlands*, pages 144–157.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, IMS Stuttgart.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Telljohann, H., Hinrichs, E., and Kübler, S. (2004). The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Tjong Kim Sang, E. F. and de Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

van Vliet, S. M. K. (2008). *Proper nouns and pronouns. The production of referential expressions in narrative discourse.* Ph.D. thesis, Tilburg University, Tilburg.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953. Association for Computational Linguistics.

Zhang, L. P. Z. and Tong, Z. (2004). Recognising and using named entities: focused named entity recognition using machine learning. In *In Proceedings of the 27th annual ACM SIGIR conference on research and development in information retrieval*, pages 281–288.