Timed Annotations — Enhancing MUC7 Metadata by the Time It Takes to Annotate Named Entities

Katrin Tomanek and **Udo Hahn**
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany
{katrin.tomanek|udo.hahn}@uni-jena.de

Abstract

We report on the re-annotation of selected types of named entities from the MUC7 corpus where our focus lies on recording the time it takes to annotate these entities given two basic annotation units – sentences *vs.* complex noun phrases. Such information may be helpful to lay the empirical foundations for the development of cost measures for annotation processes based on the investment in time for decision-making per entity mention.

1 Introduction

Manually supplied annotation metadata is at the heart of (semi)supervised machine learning techniques which have become very popular in NLP research. At their flipside, they create an enormous bottleneck because major shifts in the domain of discourse, the basic entities of interest, or the text genre often require new annotation efforts. But annotations are costly in terms of getting well-trained and intelligible human resources involved.

Surprisingly, cost awareness has not been a primary concern in most of the past linguistic annotation initiatives. Only recently, annotation strategies (such as Active Learning (Cohn et al., 1996)) which strive for minimizing the annotation load have gained increasing attention. Still, when it comes to the empirically plausible assessment of annotation costs even proponents of Active Learning make overly simplistic and empirically questionable assumptions, e.g., the uniformity of annotation costs over the number of linguistic units (e.g., tokens) to be annotated.

We here consider the time it takes to annotate a particular entity mention as a natural indicator of effort for named entity annotations. In order to lay the empirical foundations for experimentally grounded annotation cost models we couple common named entity annotation metadata with a time stamp reflecting the time measured for decision making.[1]

Previously, two studies – one dealing with POS annotation (Haertel et al., 2008), the other with named entity and relation annotation (Settles et al., 2008) – have measured the time needed to annotate sentences on small data sets and attempted to learn predictive models of annotation cost. However, these data sets do not meet our requirements as we envisage a large, coherent, and also well-known newspaper entity corpus extended by annotation costs on a fine-grained level. Especially size and coherence of such a corpus are not only essential for building accurate cost models but also as a reference baseline for cost-sensitive annotation strategies. Moreover, the annotation level for which cost information is available is crucial because document- or sentence-level data might be too coarse for several applications. Accordingly, this paper introduces MUC7$_\mathcal{T}$, our extension to the entity annotations of the MUC7 corpus (Linguistic Data Consortium, 2001) where time stamps are added to two levels of annotation granularity, *viz.* sentences and complex noun phrases.

2 Corpus Annotation

2.1 Annotation Task

Our annotation initiative constitutes an extension to the named entity annotations (ENAMEX) of the English part of the MUC7 corpus covering *New York Times* articles from 1996. ENAMEX annotations cover three types of named entities, *viz.* PERSONS, LOCATIONS, and ORGANIZATIONS. We instructed two human annotators, both advanced students of linguistics with good English language skills, to re-annotate the MUC7 corpus for the ENAMEX subtask. To be as consistent as possi-

[1] These time stamps should not be confounded with the annotation of temporal expressions (TIMEX in MUC7, or even more advanced metadata using TIMEML for the creation of the TIMEBANK (Pustejovsky et al., 2003)).

Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, pages 112–115,
Suntec, Singapore, 6-7 August 2009.

ble with the existing MUC7 annotations, the annotators had to follow the original guidelines of the MUC7 named entity task. For ease of re-annotation, we intentionally ignored temporal and number expressions (TIMEX and NUMEX).

MUC7 covers three distinct document sets for the named entity task. We used one of these sets to train the annotators and develop the annotation design, and another one for our actual annotation initiative which consists of 100 articles reporting on airplane crashes. We split lengthy documents (27 out of 100) into halves to fit on the annotation screen without the need for scrolling. Furthermore, we excluded two documents due to overlength which would have required overly many splits. Our final corpus contains 3,113 sentences (76,900 tokens) (see Section 3.1 for more details).

Time-stamped ENAMEX annotation of this corpus constitutes MUC7$_{\mathcal{T}}$, our extension of MUC7. Annotation time measurements were taken on two syntactically different *annotation units* of single documents: *(a)* complete sentences and *(b)* complex noun phrases. The annotation task was defined such as to assign an entity type label to each token of an annotation unit. Sentence-level annotation units where derived by the OPENNLP[2] sentence splitter. The use of complex noun phrases (CNPs) as an alternative annotation unit is motivated by the fact that in MUC7 the syntactic encoding of named entity mentions basically occurs through nominal phrases. CNPs were derived from the sentences' constituency structure using the OPENNLP parser (trained on PENNTREEBANK data) to determine top-level noun phrases. To avoid overly long phrases, CNPs dominating special syntactic structures, such as coordinations, appositions, or relative clauses, were split up at discriminative functional elements (e.g., a relative pronoun) and these elements were eliminated. An evaluation of our CNP extractor on ENAMEX annotations in MUC7 showed that 98.95% of all entities where completely covered by automatically identified CNPs. For the remaining 1.05% of the entity mentions, parsing errors were the most common source of incomplete coverage.

2.2 Annotation and Time Measurement

While the annotation task itself was "officially" declared to yield only annotations of named entity mentions within the different annotation units, we were primarily interested in the time needed for these annotations. For precise time measurements, single *annotation examples* were shown to the annotators, one at a time. An annotation example consists of the chosen MUC7 document with one annotation unit (sentence or CNP) selected and highlighted. Only the highlighted part of the document could be annotated and the annotators were asked to read only as much of the context surrounding the annotation unit as necessary to make a proper annotation decision. To present the annotation examples to annotators and allow for annotation without extra time overhead for the "mechanical" assignment of entity types, our annotation GUI is controlled by keyboard shortcuts. This minimizes annotation time compared to mouse-controlled annotation such that the measured time reflects only the amount of time needed for taking an annotation decision.

In order to avoid learning effects at the annotators' side on originally consecutive syntactic subunits, we randomly shuffled all annotation examples so that subsequent annotation examples were not drawn from the same document. Hence, annotation times were not biased by the order of appearance of the annotation examples.

Annotators were given blocks of either 500 CNP- or 100 sentence-level annotation examples. They were asked to annotate each block in a single run under noise-free conditions, without breaks and disruptions. They were also instructed not to annotate for too long stretches of time to avoid tiring effects making time measurements unreliable.

All documents were first annotated with respect to CNP-level examples within 2-3 weeks, with only very few hours per day of concrete annotation work. After completion of the CNP-level annotation, the same documents had to be annotated on the sentence level as well. Due to randomization and rare access to surrounding context during the CNP-level annotation, annotators credibly reported that they had indeed not remembered the sentences from the CNP-level round. Thus, the time measurements taken on the sentence level do not seem to exhibit any human memory bias.

Both annotators went through all annotation examples so that we have double annotations of the complete data set. Prior to coding, they independently got used to the annotation guidelines and were trained on several hundred examples. For the annotators' performance see Section 3.2.

[2] http://opennlp.sourceforge.net

3 Analysis

3.1 Corpus Statistics

Table 1 summarizes statistics on the time-stamped MUC7 corpus. About 60% of all tokens are covered by CNPs (45,097 out of 76,900 tokens) showing that sentences are made up from CNPs to a large extent. Still, removing the non-CNP tokens markedly reduces the amount of tokens to be considered for entity annotation. CNPs cover slightly less entities (3,937) than complete sentences (3,971), a marginal loss only.

sentences	3,113
sentence tokens	76,900
chunks	15,203
chunk tokens	45,097
entity mentions in sentences	3,971
entity mentions in CNPs	3,937
sentences with entity mentions	63%
CNPs with entity mentions	23%

Table 1: Descriptive statistics of time-stamped MUC7 corpus

On the average, sentences have a length of 24.7 tokens, while CNPs are rather short with 3.0 tokens, on the average. However, CNPs vary tremendously in their length, with the shortest ones having only one token and the longest ones (mostly due to parsing errors) spanning over 30 (and more) tokens. Figure 1 depicts the length distribution of sentences and CNPs showing that a reasonable portion of CNPs have less than five tokens, while the distribution of sentence lengths almost follows a normal distribution in the interval $[0, 50]$. While 63% of all sentences contain at least one entity mention, only 23% of CNPs contain entity mentions. These statistics show that CNPs are generally rather short and a large fraction of CNPs does not contain entity mentions at all. We may hypothesize that this observation will be reflected by annotation times.

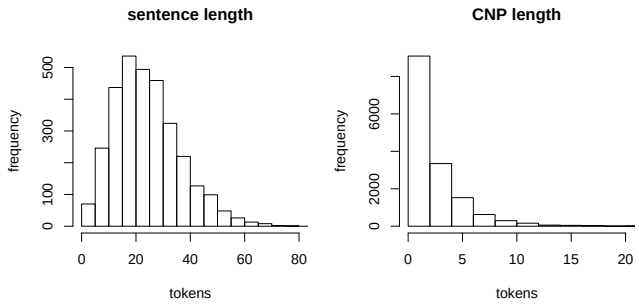

Figure 1: Length distribution of sentences and CNPs

3.2 Annotation Performance

To test the validity of the guidelines and the general performance of our annotators A and B, we compared their annotation results on 5 blocks of sentence-level annotation examples created during training. Annotation performance was measured in terms of Cohen's kappa coefficient κ on the token level and entity-segment F-score against MUC7 annotations. The annotators achieved $\kappa_A = 0.95$ and $\kappa_B = 0.96$, and $F_A = 0.92$ and $F_B = 0.94$, respectively.[3] Moreover, they exhibit an inter-annotator agreement of $\kappa_{A,B} = 0.94$ and an averaged mutual F-score of $F_{A,B} = 0.90$. These numbers reveal that the task is well-defined and the annotators have sufficiently internalized the annotation guidelines to produce valid results.

Figure 2 shows the annotators' scores against the original MUC7 annotations for the 31 blocks of sentence-level annotations (3,113 sentences) which range from $\kappa = 0.89$ to $\kappa = 0.98$. Largely, annotation performance is similar for both annotators and shows that they consistently found a block either rather hard or easy to annotate. Moreover, annotation performance seems stationary – no general trend in annotation performance over time can be observed.

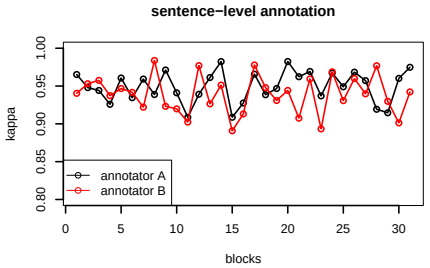

Figure 2: Average kappa coefficient per block

3.3 Time Measurements

Figure 3 shows the average annotation time per block (CNPs and sentences). Considering the CNP-level annotations, there is a learning effect for annotator B during the first 9 blocks. After that, both annotators are approximately on a par regarding the annotation time. For sentence-level annotations, both annotators again yield similar annotation times per block, without any learning effects. Similar to annotation performance,

[3] Entity-specific F-scores against MUC7 annotations for A and B are 0.90 and 0.92 for LOCATION, 0.92 and 0.93 for ORGANIZATION, and 0.96 and 0.98 for PERSON, respectively.

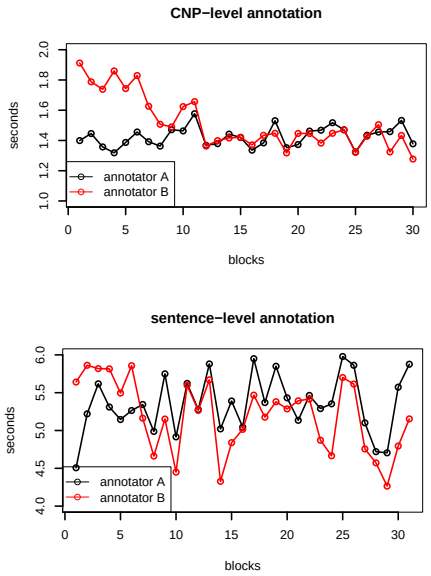

Figure 3: Average annotation times per block

analysis of annotation time shows that the annotation behavior is largely stationary (excluding first rounds of CNP-level annotation) which allows single time measurements to be interpreted independently of previous time measurements. Both, time and performance plots exhibit that there are blocks which were generally harder or easier than other ones because both annotators operated in tandem.

3.4 Easy and Hard Annotation Examples

As we have shown, inter-annotator variation of annotation performance is moderate. Intra-block performance, in contrast, is subject to high variance. Figure 4 shows the distribution of annotator A's CNP-level annotation times for block 20. A's average annotation time on this block amounts to 1.37 seconds per CNP, the shortest time being 0.54, the longest one amounting 10.2 seconds. The figure provides ample evidence for an extremely skewed time investment for coding CNPs.

A preliminary manual analysis revealed CNPs with very low annotation times are mostly short and consist of stop words and pronouns only, or

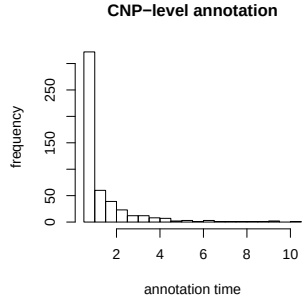

Figure 4: Distribution of annotation times in one block

are otherwise simple noun phrases with a surface structure incompatible with entity mentions (e.g., all tokens are lower-cased). Here, humans can quickly exclude the occurrence of entity mentions which results in low annotation times. CNPs which took desparately long (more than 6 seconds) were outliers indicating distraction or loss of concentration. Times between 3 and 5 seconds were basically caused by semantically complex CNPs.

4 Conclusions

We have created a time-stamped version of MUC7 entity annotations, MUC7$_\mathcal{T}$, on two levels of annotation granularity – sentences and complex noun phrases. Especially the phrase-level annotations allow for fine-grained time measurement. We will use this corpus for studies on (time) cost-sensitive Active Learning. MUC7$_\mathcal{T}$ can also be used to derive or learn accurate annotation cost models allowing to predict annotation time on new data. We are currently investigating causal factors of annotation complexity for named entity annotation on the basis of MUC7$_\mathcal{T}$.

Acknowledgements

This work was funded by the EC within the BOOTStrep (FP6-028099) and CALBC (FP7-231727) projects. We want to thank Oleg Lichtenwald (JULIE Lab) for implementing the noun phrase extractor for our experiments.

References

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

Robbie Haertel, Eric Ringger, Kevin Seppi, James Carroll, and Peter McClanahan. 2008. Assessing the costs of sampling methods in active learning for annotation. In *Proceedings of the ACL-08: HLT, Short Papers*, pages 65–68.

Linguistic Data Consortium. 2001. Message Understanding Conference 7. LDC2001T02. FTP file.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TIMEBANK corpus. In *Proceedings of the Corpus Linguistics 2003 Conference*, pages 647–656.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS'08 Workshop on Cost Sensitive Learning*, pages 1–10.