A Proposal for a Configurable Silver Standard

Udo Hahn, Katrin Tomanek, Elena Beisswanger and **Erik Faessler**

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
Fürstengraben 30, 07743 Jena, Germany
http://www.julielab.de

Abstract

Among the many proposals to promote alternatives to costly to create gold standards, just recently the idea of a fully automatically, and thus cheaply, to set up silver standard has been launched. However, the current construction policy for such a silver standard requires crucial parameters (such as similarity thresholds and agreement cut-offs) to be set *a priori*, based on extensive testing though, at corpus compile time. Accordingly, such a corpus is static, once it is released. We here propose an alternative policy where silver standards can be dynamically optimized and customized on demand (given a specific goal function) using a gold standard as an oracle.

1 Introduction

Training natural language systems which rely on (semi-)supervised machine learning algorithms, or measuring the systems' performance requires some standardized ground truth from which one can learn or against which one evaluate, respectively. Usually, a manually crafted *gold standard* is provided that is generated by human language or domain experts after lots of iterative, guideline-based training rounds. This procedure is expensive, slow and yields only small, yet highly trustable, amounts of meta data – because human experts are in the loop.

In the CALBC project,[1] an alternative approach is currently under investigation (Rebholz-Schuhmann et al., 2010a). The basic idea is to generate the much needed ground truth automatically. This is achieved by letting a flock of named entity taggers run on a corpus, without imposing any restriction on the type(s) being annotated. The (most likely) heterogeneous results are automatically homogenized subsequently, thus yielding a consensus-based, machine-generated ground truth. Considering the possible benefits (e.g., the positive experience from boosting-style machine learners (Freund, 1990)), but also being aware of the possible drawbacks (varying quality of the different systems, skewed coverage of entity types, different types of guidelines on which they were trained, etc.), the CALBC consortium refers to the outcome of this process as a *silver standard* (Rebholz-Schuhmann et al., 2010a). This procedure is inexpensive, fast, yields huge amounts of meta data – because computers are in the loop – but after all its applicability and validity has yet to be determined experimentally.

The first silver standard corpus (SSC) that came out of the CALBC project was generated by the four main partners' named entity taggers.[2] The various contributions covered, among others, annotations for genes and proteins, chemicals, diseases, etc (Rebholz-Schuhmann et al., 2010b). After the submission of their runs, the SSC was generated by, first, harmonizing stretches of text in terms of entity mention identification and, second, by mapping these normalized mentions to agreed-upon type systems (such as the MESH Semantic Groups as described by Bodenreider and McCray (2003) for entity type normalization). Basically, the harmonization steps included rules when entity mentions were considered to match or overlap (using a cosine-based similarity criterion) and entity types referred to the same class. For consensus generation, finally, simple rules for majority votes were established.

The CALBC consortium is fully aware of the fact that the value of an SSC can only be assessed

[1] http://www.calbc.eu

[2] The CALBC consortium consists the Rebholz Group from EBI (Hinxton, U.K.), the Biosemantics Group from Erasmus (Rotterdam, The Netherlands), the JULIE Lab (Jena, Germany), and LINGUAMATICS (Cambridge, U.K.).

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 235–242,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

by comparing, e.g., systems trained on such a silver standard with systems trained on a gold standard (preferably, though not necessarily, one that is a subset of the document set which makes up the SSC).

In the absence of such a gold standard, the CALBC consortium has spent enormous efforts to find out the most reasonable parameter settings for, e.g., the cosine threshold (setting similar mentions apart from dissimilar ones) or the consensus constraint (where a certain number of entity types equally assigned by different taggers makes one type the consensual silver one and discards all alternative annotations). Once these criteria are made effective, the SSC is completely fixed.

As an alternative, we are looking for a more flexible solution. Our investigation was fuelled by the following observations:

- The idiosyncrasies of guidelines (on which (some) taggers were trained) do not necessarily lead to semantically totally different entities although they differ literally to some degree. Some guidelines prefer, e.g., *"human IL-7 protein"*, others favor *"human IL-7"*, and some lean towards *"IL-7"*. As the cosine measure tends to penalize a pair such as *"human IL-7 protein"* and *"IL-7"*, we intended to avoid such a prescriptive mode and just look at the type assignment for single tokens as (parts of) entity mentions. thus avoiding inconclusive mention boundary discussions.

- While we were counting, for all tokens of the document set, the votes a single token received from different taggers in terms of annotating this token with respect to some type, we generated confidence data for meta data assignments. Incorporating the distribution of confidence values into the configuration process, this allows us to get rid of *a priori* fixed majority criteria (e.g., two or three out of five systems must agree on this token) which are hard to justify in an absolute way.

Summarizing, we believe that the nature of diverging tasks to be solved, the levels of entity type specificity to be reached, the sort of guidelines being preferred, etc. should allow prospective users of a silver standard to *customize* one on their own and not stick to one that is already prefabricated without concrete application in mind.[3]

As such an enterprise would be quite arbitrary without a reference standard, we even go one step further. We determine the suitability of, say, different voting scores and varying lexical extensions of mentions by comparison to a gold standard so that the 'optimal' configuration of a silver standard, given a set of goal-derived requirements, can be automatically learned. In real-world applications, such gold standard annotations would be delivered only for a fraction of the documents contained in the entire corpus being tagged by a flock of taggers. The gold standard is used to optimize parameters which are subsequently applied to the aggregation of automatically annotated data. Note that the gold standard is used for optimization only, not for training. We call such a flexible, dynamically adjustable silver standard a *configurable Silver Standard Corpus* (conSSC). In a second step, we split the various conSSCs, re-trained our NER tagger on these data sets and, by comparison with the gold standard, were able to identify the optimal conSSC for this task (which is not the one (SSC I) made available by the CALBC consortium for the first challenge round).[4]

2 Optimizing Silver Standards

In this section, we describe the constituent parameters of a wide spectrum of SSCs. Mostly, these parameters were taken over from the design of the SSC as developed by the CALBC project members. Differing from that fixed SSC, we investigate the impact of different parameter settings on the construction of a collection of SSCs, and, first, evaluate their direct usefulness on a gold standard for protein-gene annotations. Second, we also assess their indirect usefulness by training NER classifiers on these SSCs and evaluate the NERs' performance on the gold standard. Thus, our approach is entirely data-driven without the need for human intervention in terms of choosing suitable parameter settings.

Technically, we first aggregate the votes from the flock of taggers (in our experiments, we used the four taggers from the CALBC project members plus a second tagger of one of the members) for each text token (for confidence-based decisions) or at the entity level (for cosine-based decisions), then we determine the confidence values of these

[3] There may be tasks where a "long" entity such as *"human IL-7 protein"* may be appropriate, while for another task a short one such as *"IL-7"* is entirely sufficient.

[4] http://www.ebi.ac.uk/Rebholz-srv/CALBC/challenge.html

aggregated votes, and, finally, we compute the similarity of the various SSCs with the gold standard data in terms of F-scores (both exact and open boundaries) and accuracy on the token level.

2.1 Calibrating Consensus

The metrical interpretation of consensus will be based on thresholded votes for semantic groups at the token level (cf. Section 2.1.1) and a cosine-based measure to determine contiguous stretches of entity mentions in the text (cf. Section 2.1.2).

2.1.1 Type Confidence and Type Voting

For each text token, we determine the entity type assignment as generated by each NER tagger which is part of the flock of CALBC taggers.[5] We count and aggregate these votes such that each entity type has an associated type count value.

We then compute the ratio of systems agreeing on the same single type assignment and call this the *confidence* attributed to a particular type for some token. The confidence value will subsequently be interpreted against the *confidence threshold* $[0, 1]$ that defines a measure of certainty a type assignment should have in order to be accepted as consensual.

2.1.2 Cosine-based Similarity of Phrasal Entity Mentions

As the above policy of token-wise annotation decouples contiguous entity mentions spanning over more than one token, we also want to restitute this phrasal structure. This is achieved by constructing contiguous sequences of tokens that characterize a phrasal entity mention at the text level to which the same type label has been assigned. Since different taggers tend to identify different spans of text for the same entity type (as shown in the example from Section 1) we have to account for similar phrasal forms of named entity mentions.

This is achieved by constructing vectors which represent entity mentions and by computing the cosine between the different entity mention vectors. Let $E_1 = T_1T_2T_3$ be an entity mention comprised of three tokens T_1 to T_3. Let $E_2 = T_2T_3$ be an entity mention overlapping with E_1 in the tokens T_2 and T_3. To decide whether E_1 and E_2 are considered similar, we first construct two vectors representing the entity mentions:

$$v(E_1) = (f_1, f_2, f_3)^T$$

with $f_i = IDF(T_i)$ being the inverse document frequency of the token T_i. We compute the inverse document frequency of tokens based on the corpus which is subject to analysis. Analogously, we construct the vector for E_2

$$v(E_2) = (0, f_2, f_3)^T$$

filling in a zero for the IDF of T_1 since it is not covered by E_2. The entity mentions E_1 and E_2 are considered equal or similar, if the cosine of the two vectors is greater or equal a given threshold, $\cos(v(E_1), v(E_2)) \geq threshold$.[6] We then compute the number of systems considering an entity annotation as similar in the manner described above. The annotation is accepted and thus entered into the SSC, if a particular number of systems agree on one annotation. This approach was previously developed by the CALBC project partners (Rebholz-Schuhmann et al., 2010a).

The number of agreeing systems and the threshold are the free parameters of this method and thus subject to optimization.

2.2 Optimization of Silver Standard Corpora

In the experiments described in the next section, we will consider alternative parametrizations for Silver Standard Corpora, i.e., the required confidence threshold or cosine threshold and the number of agreeing systems. We will then discuss two variants for optimizing this collection of SSCs. The first one directly uses the gold standard for optimization. The task will be to find that particular parameter setting for an SSC which best fits the data contained in the gold standard. Once these parameters are determined they can be applied to the complete CALBC document set (composed of 100,000 documents) to produce the final, quasi-optimal SSC.

In another variant, we insert a classifier into this loop. First, we train a classifier on a particular

[5] Due to time constraints when we performed our experiments, we make an extremely simplifying assumption: From the whole range of possible entity types NER taggers may assign to some token (cf. (Bodenreider and McCray, 2003)) we have chosen the PRotein/GEne group for testing. Still, this assumption does not do harm to the core of our hypotheses. See also our discussion in Section 5.

[6] For final corpus creation, it must be decided which of the matching entity mentions is entered into the reference SSC, e.g. the longest or shortest entity annotation. In our experiments, we always chose the shortest entity mention. However, preliminary experiments showed that the differences to taking the longest entity mention were marginal.

SSC that is built from a particular parameter combination. Next, this classifier is tested against the gold standard. This is iterated through all parameter combinations. Obviously, the best performing classifier relative to the gold standard selects the optimal SSC.

3 Experimental Setting

3.1 Gold Standard

We generated a new broad-coverage corpus composed of 3,236 MEDLINE abstracts (35,519 sentences or 941,890 tokens) dealing with gene and protein mentions. Altogether, it comprises 57,889 named entity type annotations annotated by one expert biologist. We created this new resource to have a consistent and (as far as possible) subdomain-independent protein-annotated corpus.[7]

MEDLINE abstracts were annotated with (protein coding) genes, mRNAs and proteins. A distinction was made between dedicated proteins as they are recorded in the protein database UNIPROT,[8] protein complexes consisting of several protein subunits (e.g., IL-2 receptor consisting of α, β, and γ chain), and protein families or groups (e.g., "transcription factors"). Also enumerations of proteins and protein variants were annotated. Discontinuous annotations were avoided as well as nested annotations (annotations embedded in other annotations). However, gene/protein mentions nested in terms other than gene/protein mentions were annotated (e.g., protein mentions nested in protein function descriptions such as *"ligase"* in *"ligase activity"*). Modifiers such as species designators were excluded from annotations whenever possible. Gene segments or protein fragments were also not annotated.

For our experiments, we did not distinguish between the different annotation classes (see Table 1) but merged all available annotations into one class, *viz.* PRotein/GEne (PRGE).

3.2 Automatic Annotation of the Gold Standard

We then asked all four sites participating in the CALBC project to automatically annotate the given gold standard (made available without gold data, of course) using the same type of named entity tagging machinery as was used to annotate CALBC's canonical SSC. The performance results of each group's system evaluated against the gold standard are reported in Table 2. The data of each system constitute the reference data sets and raw data for all subsequent experiments on the configuration and optimization of the silver standard.

semantic type	description
T028	Gene or Genome
T086	Nucleotide Sequence
T087	Amino Acid Sequence, Amino Acid, Peptide
T116	Protein
T126	Enzyme
T192	Receptor

Table 1: Semantic types defining the PRGE group (semantic type codes refer to the UMLS).

The resulting raw material does thus not only contain gene/protein annotations but also any other entity types as supplied by the partners. For our experiments on the gold standard, however, only the entity types subsumed by the PRGE group (see Table 1) were considered and annotations of all other types were discarded. The definition of the PRGE group is identical to the one proposed by Rebholz-Schuhmann et al. (2010a). For the experiments, the specific semantic types (e.g., the UMLS concepts)[9] were not considered, only the semantic group PRGE was.

3.3 Evaluation Metrics

The following metrics were used to evaluate how good the silver standard(s) fit(s) the provided gold standard:

- segment-level recall, precision, and F-score values with exact boundaries, the standard way to evaluate NER taggers,
- segment-level recall, precision, and F-score, but with relaxed boundary constraints. This means that two entity mentions are considered to match when they overlap with at least one token and have the same entity type assigned to them,
- accuracy measured on the token level.

These metrics can be considered as optimization criteria.

[7] We are aware of other gene/protein-annotated corpora such as PENNBIOIE (`http://bioie.ldc.upenn.edu/`) or GENIA (`http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi`) that will have to be taken into account in future studies as well.

[8] `http://www.uniprot.org/`

[9] `http://www.nlm.nih.gov/research/umls/`

3.4 Tokenization

The CALBC partners' data do not necessarily come with tokenization information and, moreover, different partners/systems might have different tokenizations. Since a common ground for comparison is thus lacking we added a new, consistent tokenization based on the JULIE Lab tokenizer (Tomanek et al., 2007b). This tokenizer is optimized for biomedical documents with intrinsic focus to keep complex biological terminological units (such as *"IL-2"*) unsegmented, but to split up tokens that are not terminologically connected (such as dividing *"IL-2-related"* up into *"IL-2"*, *"-"* and *"related"*). As a matter of fact, entity boundaries do not necessarily coincide with token boundaries. Our solution to this problem is as follows: Whenever a token partially overlaps with an entity name, the full form of that token is considered to be associated with this entity. All data on which we report here (silver and gold standards) obey to this tokenization scheme.

3.5 Parameters Being Tested

The following parameter settings were considered in our experiments:

- Four different values for confidence thresholds indicating that 20% (0.2), 40% (0.4), 60% (0.6) or 80% (0.8) of all taggers agreed on the same type annotation, *viz.* PRGE,
- Five different values for cosine thresholds to identify overlapping entity mentions, *viz.* $(0.7, 0.8, 0.9, 0.95, 0.975)$, and two different values for the number n of agreeing taggers, *viz.* $n \geq 2$ and $n \geq 3$,
- Two tagger crowd scenarios, *viz.* one where *all five* systems were involved, the other where subsets of cardinality *2* of these crowds were re-combined.[10]

4 Results

As already described in Section 2.2, we performed two types of experiments. In the first experiment (Section 4.1), we intend to find proper calibrations of parameters for an optimal SSC as described in Section 3.5. In the second experiment (Section 4.2), we incorporate an extrinsic task, training an NER classifier on different parameter settings, as a selector for the optimal SSC.

[10] We refrained from also testing combinations of *3* and *4* systems due to time constraints.

4.1 Intrinsic Calibration of Parameters

Full Merger of All Taggers. In this scenario, we tested the merged results of the entire crowd of CALBC taggers when compared to the gold standard and determined their performance scores (see Table 3). We will discuss the results with respect to the overlapping F-score, if not explicitly stated otherwise.

Looking at the results of the runs involving different *cosine* thresholds, we witness a systematic drawback when more than two systems are required to agree. Although precision is boosted in this setting, recall is decreasing strongly which results in overall lower F-scores. When only two systems are required to agree a comparatively higher recall comes at the cost of lower precision. Yet, the F-score (both under exact as well as overlap conditions) is always superior (ranging between 75% and 73%) when compared to the 3-agreement scenario. Note that the 2-agreement condition for the highest threshold being tested yields, without exception, better scores than the best single system (cf. Table 2).

The best performing run in terms of F-score for the *confidence* method results from a threshold of 0.2 with an F-score of 76%. Note that this F-score lies *4* percentage points above the best performance of a single system (cf. Table 2).

A threshold of *0.2* with five contributing systems results in a union of all annotations. Consequently, this run benefits from a high recall compared with the other runs. However, the run exhibits the lowest precision rating (both for the exact and overlap condition), which is due to the low threshold being chosen. As can also be seen with the confidence method at a threshold of 0.80, a very high precision can be reached (99%) but at the cost of extremely low recall.[11] The methods performing best in terms of overlapping F-score also perform best in terms of exact F-score.

Selected Tagger Combinations: Twin Taggers. In this scenario, we evaluated all twin combinations of taggers against the gold standard regarding the confidence criterion. In Table 4 we contrast the two best performing and the two worst performing tagger pairs for the confidence method. The table reveals that there are some cases where the taggers seem to complement each other, e.g., the twins SYS-1 and SYS-3, as well as SYS-3 and

[11] Exactly these kinds of alternatives offer flexibility for choosing the most appropriate SSC given a specific task.

exactR	exactP	exactF	overlapR	overlapP	overlapF	systems
0.55	0.74	**0.63**	**0.63**	0.84	**0.72**	SYS-1
0.36	0.53	0.43	0.46	0.68	0.55	SYS-2
0.48	0.77	0.59	0.59	**0.95**	**0.72**	SYS-3
0.44	**0.83**	0.58	0.49	0.91	0.64	SYS-4
0.34	0.61	0.44	0.41	0.74	0.53	SYS-5

Table 2: Performance of single systems (SYS-1 to SYS-5) as evaluated against the gold standard (best performance scores in bold face). Measurements are taken both for exact as well as overlapping recall (R), precision (P) and F-score (F).

method	ACC	exactR	exactP	exactF	overlapR	overlapP	overlapF	threshold	agr. systems
cosine	**0.94**	**0.53**	0.71	**0.61**	**0.66**	0.87	**0.75**	0.70	2.00
cosine	0.93	0.40	**0.79**	0.53	0.49	**0.96**	0.65	0.70	3.00
cosine	**0.94**	**0.54**	0.71	**0.61**	**0.65**	0.87	**0.74**	0.80	2.00
cosine	0.93	0.41	**0.80**	0.54	0.48	**0.95**	0.64	0.80	3.00
cosine	**0.94**	**0.54**	0.72	**0.62**	**0.65**	0.86	**0.74**	0.90	2.00
cosine	0.93	0.41	**0.81**	0.54	0.48	**0.95**	0.64	0.90	3.00
cosine	**0.94**	**0.54**	0.73	**0.62**	**0.64**	0.86	**0.74**	0.95	2.00
cosine	0.93	0.41	**0.83**	0.55	0.47	**0.95**	0.63	0.95	3.00
cosine	**0.94**	**0.55**	0.75	**0.64**	**0.64**	0.86	**0.73**	0.97	2.00
cosine	0.93	0.42	**0.85**	0.56	0.47	**0.95**	0.63	0.97	3.00
confidence	**0.95**	**0.58**	0.73	**0.65**	**0.68**	0.85	**0.76**	0.20	
confidence	0.94	0.44	0.83	0.58	0.50	0.94	0.66	0.40	
confidence	0.93	0.32	0.88	0.47	0.35	0.97	0.52	0.60	
confidence	0.91	0.16	**0.91**	0.27	0.17	**0.99**	0.30	0.80	

Table 3: Merged annotations of the entire crowd of CALBC taggers (best performance scores per parameter setting in bold face). Parameters: threshold (confidence or cosine) and number of agreeing systems (agr. systems).

SYS-4. In both cases, a confidence threshold of 0.2 yields the best F-score. Additionally, these F-scores (81% and 78%) are even higher than the single system's F-scores ($+9\%$ up to $+14\%$). This comes with a significant increase in recall over both systems ($+13\%$ to $+28\%$) though at the cost of lowered precision relative to the system with the higher precision (-1% to -10%). These results also outperform the best results of the experimental runs where all systems were involved (see Table 3). This indicates that a subset of all systems might yield a better SSC than a combination of all systems' outputs.

4.2 Extrinsic Calibration of Parameters

We employed a standard named entity tagger to assess the impact of the different merging strategies on a scenario near to a real-world application.[12] Each SSC variant (and thus each parameter combination) was evaluated with this tagger in a 10-fold cross validation. The SSC and the gold corpus were split into ten parts of equal size. Nine parts of the SSC constituted the training data of one cross validation round, the corresponding tenth part of the gold standard was used for evaluation. This way, we tested how adequate a merged corpus was with respect to the training of a classifier. Because the cross validation has been very time consuming, we did not consider specific combinations of systems but always merged the annotations of all five systems. The results are displayed in Table 5.

Interestingly, the highest recall, precision, and F-score values (both for the exact and overlap condition) are shared by the same parameter combinations which also performed best in Section 4.1. Hence, the use of a named entity tagger supports the evaluation results when comparing the various

[12] This tagger is based on Conditional Random Fields (Lafferty et al., 2001) and employs a standard feature set used for biomedical entity recognition (Settles, 2004).

ACC	exactR	exactP	exactF	overlapR	overlapP	overlapF	systems	threshold
0.95	0.62	0.69	0.65	**0.76**	**0.85**	**0.81**	**SYS-1 + SYS-3**	0.20
0.92	0.22	0.69	0.34	0.26	0.81	0.39	SYS-2 + SYS-5	0.60
0.95	0.55	0.75	0.63	**0.67**	**0.91**	**0.78**	**SYS-3 + SYS-4**	0.20
0.92	0.30	0.85	0.45	0.34	0.94	0.50	SYS-4 + SYS-5	0.60

Table 4: Twin pairs of taggers, contrasting the two best (in bold face) and the two worst performing pairs obtained by the confidence method.

method	ACC	exactR	exactP	exactF	overlapR	overlapP	overlapF	threshold	agr. systems
cosine	**0.94**	**0.46**	0.69	**0.56**	**0.58**	0.86	**0.69**	0.70	2.00
cosine	0.93	0.32	**0.77**	0.45	0.39	**0.94**	0.55	0.70	3.00
cosine	**0.94**	**0.46**	0.69	**0.56**	**0.57**	0.86	**0.69**	0.80	2.00
cosine	0.93	0.32	**0.78**	0.46	0.39	**0.94**	0.55	0.80	3.00
cosine	**0.94**	**0.46**	0.70	**0.56**	**0.57**	0.85	**0.68**	0.90	2.00
cosine	0.93	0.32	**0.79**	0.46	0.38	**0.93**	0.54	0.90	3.00
cosine	**0.94**	**0.47**	0.71	**0.56**	**0.56**	0.85	**0.68**	0.95	2.00
cosine	0.93	0.33	**0.80**	0.47	0.38	**0.93**	0.54	0.95	3.00
cosine	**0.94**	**0.47**	0.73	**0.57**	**0.56**	0.85	**0.67**	0.97	2.00
cosine	0.93	0.33	**0.82**	0.47	0.38	**0.93**	0.54	0.97	3.00
confidence	**0.94**	**0.50**	0.72	**0.59**	**0.60**	0.85	**0.70**	0.20	
confidence	0.93	0.36	0.82	0.50	0.41	0.93	0.56	0.40	
confidence	0.92	0.25	0.87	0.39	0.28	0.95	0.43	0.60	
confidence	0.91	0.12	**0.89**	0.20	0.12	**0.96**	0.22	0.80	

Table 5: Performance of an NER tagger trained on an SSC, 10-fold cross validation, and all systems. Parameters: threshold (confidence or cosine) and number of agreeing systems (agr. systems).

SSCs directly to the gold standard corpus. However, this result may be due to our particular experimental setting and should not be taken as a general rule. Instead, this issue should be studied on additional gold standard corpora (cf. Section 5).

5 Discussion and Conclusions

The experiments reported in this paper strengthen the empirical basis of the novel idea of a silver standard corpus (SSC). While the originators of the SSC have come up with a fixed SSC, our experiments show that different parametrizations of SSCs allow to dynamically configure or select an optimal one given a gold standard for comparison during this optimization.

Our experimental data reveals that the boosting hypothesis (the combination of several classifiers outperforms weaker single ones in terms of performance) is confirmed for complete mergers as well as selected twin pairs of taggers. We also have evidence that boosting within the SSC paradigm tends to increase precision whereas it seems to decrease recall. This general observation becomes stronger and stronger when the size of the committees (i.e., the number of submitting classifiers) increases. It is also particularly interesting that both the intrinsic evaluation (groups of classifiers *vs.* gold standard), as well as the extrinsic evaluation of SSCs (groups of classifiers trained and tested on mutually exclusive partitions of the gold standard) reveal parallel patterns in terms of performance – this indicates a surprising level of stability of the entire SSC approach.

In our view, the strongest finding from our experiments is the possibility to calibrate an SSC according to requirements derived from the goal of annotation campaigns. In particular, one can adapt parameters to a specific use case, e.g., building a corpus with high precision when compared to the gold standard. Through the evaluation of the parameter space, one can assess the costs of reaching a specific goal. For instance, a precision of 99% can be reached, yet at the cost of the F-score plunging to 30%; only slightly lowering the precision to 97% boosts the F-score by 22 points (see last two rows in Table 3).

Also, when increasingly more annotation sets become available (e.g., through the CALBC challenges) the problem of adversarial or extremely bad performing systems is no longer a pressing issue since with the optimization approach such systems are automatically sorted out when optimizing over the set of possible system combinations.

While our experiments are but a first step towards the consolidation of the SSC paradigm some obvious limitations of our work have to be overcome:

- experiments with different gold standards have to be run as one might hypothesize that different gold standards require different parameter settings for the optimal SSC,
- experiments with different NER taggers have to be run (e.g., we plan to use an NER tagger which prefers recall over precision, while the one used for these experiments generally yields higher precision than recall scores),
- test with crowds of taggers which generate higher recall than precision.[13]

In our approach, a gold standard is needed to find good parameters to build an SSC. A question not addressed so far is how huge such a gold standard must be to offer an appropriate size for the optimization step. Finally, it might be particularly rewarding to join efforts in reducing the development costs for such a gold standards – Active Learning (e.g., Tomanek et al. (2007a)) might be one promising approach to break this bottleneck. Since effective calibration of SSCs is in need of reasonably sized and densely populated gold standards, by combining these lines of research we claim that additional benefits for SSCs become viable.

6 Acknowledgments

We wish to thank Kerstin Hornbostel for stimulating and corrective remarks on the biological grounding of this investigation. This research was partially funded by the EC's 7th Framework Programme within the CALBC project (FP7-231727) and the GERONTOSYS research initiative from the German Federal Ministry of Education and Research (BMBF) under grant 0315581D within the JENAGE project.

[13] We used a gold standard in which some unusual entities (e.g., protein families) had been annotated for which most named entity taggers have not been trained. This might also explain the generally overall low recall among the crowd of taggers yielded in our experiments.

References

Olivier Bodenreider and Alexa T. McCray. 2003. Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36(6):414–432.

Yoav Freund. 1990. Boosting a weak learning algorithm by majority. In *COLT'90 – Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pages 202–216.

John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01 – Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010a. CALBC Silver Standard Corpus. *Journal of Bioinformatics and Computational Biology*, 8:163–179.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, Peter Milward, David Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, and Udo Hahn. 2010b. The CALBC Silver Standard Corpus for biomedical named entities: A study in harmonizing the contributions from four independent named entity taggers. In *LREC 2010 – Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *NLPBA/BioNLP 2004 – COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 107–110.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007a. An approach to text corpus construction which cuts annotation costs and maintains corpus reusability of annotated data. In *EMNLP-CoNLL'07 – Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning*, pages 486–495.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007b. A reappraisal of sentence and token splitting for life sciences documents. In K. A. Kuhn, J. R. Warren, and T. Y. Leong, editors, *MEDINFO'07 – Proceedings of the 12th World Congress on Medical Informatics*, number 129 in Studies in Health Technology and Informatics, pages 524–528. IOS Press.