Centroids: Gold standards with distributional variations

Ian Lewin*, Şenay Kafkas†, Dietrich Rebholz-Schuhmann†

Linguamatics*
St. John's Innovation Centre
Cowley Road
Cambridge
UK

European BioInformatics Institute†
Wellcome Trust Genome Campus
Hinxton
Cambridge
UK

ian.lewin@linguamatics.com, {kafkas,rebholz}@ebi.ac.uk

Abstract

Motivation: Gold Standards for named entities are, ironically, not standard themselves. Some specify the "one perfect annotation". Others specify "perfectly good alternatives". The concept of Silver standard is relatively new. The objective is consensus rather than perfection. How should the two concepts be best represented and related? Approach: We examine several Biomedical Gold Standards and motivate a new representational format, centroids, which simply and effectively represents name distributions. We define an algorithm for finding centroids, given a set of alternative input annotations and we test the outputs quantitatively and qualitatively. We also define a metric of relatively acceptability on top of the centroid standard. Results: Precision, recall and F-scores of over 0.99 are achieved for the simple sanity check of giving the algorithm Gold Standard inputs. Qualitative analysis of the differences very often reveals errors and incompleteness in the original Gold Standard. Given automatically generated annotations, the centroids effectively represent the range of those contributions and the quality of the centroid annotations is highly competitive with the best of the contributors. Conclusion: Centroids cleanly represent alternative name variations for Silver and Gold Standards. A centroid Silver Standard is derived just like a Gold Standard, only from imperfect inputs.

Keywords: Centroid, Gold Standard, Silver Standard, Evaluation

1. Introduction

We examine several Gold Standard assessment datasets available in biomedicine and motivate a new representation for Gold Standard markup: *centroids*. Centroids provide a simple and effective representation for name *distributions* and a more fine-grained method for measuring how good a user annotation is. In this way, centroids represent an *extension* of classical gold standard markup.

In addition, we define an algorithm for finding centroids, given a set of alternatively annotated inputs, and test it quantitatively and qualitatively against both Gold Standard inputs and automatically annotated inputs.

Given a set of alternative inputs, each of which is Gold Standard, we verify that the algorithmically discovered centroids are also overwhelmingly gold standard, as traditionally conceived. Even when (infrequently) they are not, they very often represent errors in the original gold standard.

We apply the algorithm also to sets of alternative automatic annotations as submitted to the CALBC challenge competition. We thereby derive a *Silver* Standard, a representation of a consensus driven standard. We show that silver standard centroids are very highly competitive with the best of the contributing annotations. Further experiments also show that Silver Standard scores correlate to Gold Standard, suggesting that silver might indeed stand proxy for gold as well as representing a consensus annotation (with distributions).

We conclude that it is indeed highly desirable to represent distributions directly within the Gold Standards and not just implicitly, for example through fuzzy or partial matching schemes.

2. Background

The Gold Standard data-sets available for biomedical named entity recognition are only few in number. Yet, they differ in more than just subject matter. Here we show how they vary in the representation and suggested evaluation of name variations.

The SCAI corpus (Kolarik et al., 2008) for chemical entities, for instance, assigns gold standard I-O-B labels to predefined *tokens* within sentences. Good scores require reproducing the same tokenization as well as the one correct label for each token. Some variation is possible through the use of special class labels. For example, the token *compounds* within *uridine compounds* is assigned the label `B-Modifier`. This perhaps indicates that *compounds* is not here functionning as an essential part of the name of the chemical. Other labels, such as `B-Trivial` and `B-Family` are similarly suggestive. It is therefore at least possible for an evaluation procedure to be sensitive to these

class-labels and treat the actual tags generated by an automatic annotator accordingly.

A second example is BioCreative (Krallinger et al., 2008). Here, gold standard entities are defined over untokenized character strings through the use of character offsets. The offsets mean that automatic annotators need not re-use a predefined tokenization. They also easily allow for the markup of alternative names. In a sentence 20 characters long, one entity may be defined from character positions 10 to 12 whereas another may be defined from 7 to 15. These entities overlap and represent alternative markups. Clearly, it is also possible for two names that do not overlap to both overlap a third.

In BioCreative, one 'core' set of entirely non-overlapping offsets is dignified with the title "Gene List." All other offsets are included in a second and optional list called the "Alternative Gene List." Fig 1 illustrates one example where the Gene List entry *endogenous TGF-beta-specific complex* overlaps three smaller entries in the Alternative Gene List.

BioCreative supplies its own evaluation procedure which, if the Alternative Gene List is left unspecified, sets automatic annotators the task of reproducing exactly all and only the names in the 'core' Gene List. If the Alternative Gene List is also specified, then the task becomes slightly more complex. The intuition behind it is that a candidate annotation is true positive simply if it appears either in the Gene List or in the Alternative Gene List. Also, a false negative only occurs if a *core name* (from the Gene List) is absent from the annotation and every alternative (in the Alternative Gene List) that overlaps it is also absent[1]. Therefore perfect recall requires spotting at least one alternative for every core name. It should be noted that all alternatives are deemed equally as good as each other and equally as good as the "core" name that they overlap.

Finally, we briefly consider the Arizona Disease Corpus (Leaman et al., 2009). This also uses offsets and includes some alternative overlapping names. Unlike BioCreative however, all alternatives are included within one flat list and no evaluation script is provided so it is somewhat unclear how the alternatives are to be scored. Fig 2 illustrates some example alternative names from the Arizona Disease Corpus.

In what follows, we propose a new representation in which the *core* name is identified (automatically) and the alternatives are represented as a distribution around it. We also define an intuitive evaluation scheme for candidate markup against such a gold standard and report on a number of experiments both with gold standard inputs and automatic inputs.

3. Centred Alternatives & Preferred Alternatives

The BioCreative names in Fig. 1 all share a common heart: *TGF-beta*. Every name is a string extension of one string: *TGF-beta*, and that string is itself Gold Standard. Let us call this property the *centroid* property. One might expect *all* sets of alternatives to exhibit it, but, as we show below, this is not actually true of all the Gold Standards we have examined. By contrast, we suggest a representation that *requires* it.

[1]There are small wrinkles which disturb this intuitive picture. They are discussed further below.

Core Name	endogenous TGF-beta-specific complex
Alternatives	TGF-beta endogenous TGF-beta TGF-beta-specific complex
Sentence	... similar to that of the endogenous TGF-beta-specific complex observed in ...

Figure 1: BioCreative alternatives

Sentence	... predispose carriers to multiple adenomatous polyps of the colon and rectum and to ...
Names	multiple adenomatous polyps adenomatous polyps of the colon adenomatous polyps of the colon and rectum

Figure 2: AZDC alternatives (no names are identified as "core")

Certainly, the centroid property does not hold true of the AZDC names in Fig. 2. A common heart exists, *adenomatous polyps* but it is not itself deemed to be a Gold Standard name. It is very unclear what the reason for this might be. The internal evidence is highly suggestive: the last two names suggest that *multiple* is optional; the first suggests *of the colon* is optional. Externally, one easily verifies that *Adenomatous polyps* lies within the Mesh controlled vocabulary (id:D018256).

The centroid property indicates that alternative names should have a Gold Standard common heart. In addition however, we may observe that the alternative names in Fig 1 have a distribution. Of all the extensions of *TGF-beta*, two include *endogenous* and two do not. Similarly, two names post-fix *-specific complex* and two do not. However, no names only post-fix *-specific*.

In our proposals below, we both represent these facts and we use them for judgments of *relative acceptability*. The measure we define below (section 6.) has the result that an annotation that include *endogenous* is just as preferred as one that omits it. The measure makes a boundary after *-specific* less acceptable than a boundary after *TGF-beta*, whereas the latter boundary is just as acceptable as one after *complex*. In this way, many expert (or indeed automatic) judgments can help contribute to a distributional assessment of correctness.

4. Centroid Algorithm

Given a corpus of text, the input to the centroid algorithm is a set of markups over that corpus. Each markup consists of the text considered as a character string plus inline markup of (non-overlapping) entity names. Given the Context and Names shown in Fig 1, four inline markups would be required, each containing one of the alternative names.

Texts are tokenized at the character level and (ignoring spaces) votes are counted over pairs of *adjacent name-*

```
... o-f-t-h-e-e-n-d-o-g-e-n-o-u-s-T-G-F-b-e-t-a-s-p-e-c-i-f-i-c-c-o ...
...  0 0 0 0 0 2 2 2 2 2 2 2 2 2 2 2 4 4 4 4 4 4 2 2 2 2 2 2 2 2 2 2 ...
```

Figure 3: Counts of adjacent name-internal characters

1) ... of the endogenous ⟨e b='l:0:2,l:10:2,r:0:2,r:16:2'⟩TGF-beta⟨/e⟩-specific complex observed ...
2) ... Rel-related human ⟨e b='l:16:2,l:5:3,l:0:2,r:0:2,r:13:1,r:20:3'⟩p75⟨/e⟩nucleoprotein complex ...

Figure 4: Two centroids + boundary distributions

internal characters. Figure 3 shows the character-pair counts for our running example. Two markups consider that the transition $e \rightarrow n$ at the beginning of *endogenous* falls within the name; whereas all four consider that $T \rightarrow G$ do. The focus on name-internal pairs, rather than single characters, ensures that boundaries are valued when two different names happen to be directly adjacent themselves.

Over the course of a text, the number of votes will mostly be zero, punctuated by occasional bursts of wave-like variation. We define the centroids to be the substring(s) over character pairs that are peaks (or local maxima) in a burst of votes. In our running example, there is only one peak and the centroid is *TGF-beta*.

In addition, thresholds may be added to further refine the notion of a peak. For example, "only local maxima" might be relaxed to "local maxima plus surrounding material within a certain threshold of votes". In this way, small deviations from peaks might be ignored. Equally, a minimum threshold of votes may be applied so that very low peaks become ignored. Thresholding becomes important when the algorithm is applied to generate a Silver Standard from sets of imperfect inputs, such as those that result from automatic annotation.

The character-pair votes also define the boundary distribution around the centroid. We currently define a possible boundary whenever the number of votes changes. Its value is the difference in votes. So, the centroid *TGF-beta* has a possible left boundary before *T* and its value is 2 (the difference between 4 and 2). The boundary before *endogenous* also scores 2 (the difference between 0 and 2). There is *no* boundary immediately after *specific*.

Fig. 4 illustrates typical outputs of the algorithm using an xml representation. The centroid itself is represented inside an e element. Each possible boundary is represented as `d:p:v` (direction,position,value) in a comma-delimited list. For example, `l:10:2` means that there is a boundary 10 characters to the left of *TGF-beta*. This boundary includes *emdogenous*. The 'value' of the boundary, 2, is used in measuring relative acceptability (see further below). The string encompassed by the leftmost and rightmost boundaries is the *maximal extent* of the centroid.

5. Algorithm Evaluation

Here we evaluate the centroid algorithm first when fed perfect inputs, namely sets of humanly annotated alternatives from Gold Standard corpora, and secondly when fed automatic annotations.

5.1. Gold Standard Inputs

We generated multiple copies of the BioCreative training corpus (15k sentences) such that every Gold Standard name (core and alternative) was marked up once in some copy of the corpus. Then, we applied the centroid algorithm. No thresholding was used. As the inputs are all Gold Standard, no threshold ought to be necessary because one vote is simply enough votes.

We evaluated the resulting centroids using, as a de facto pre-existing standard, the BioCreative evaluation script. (For the use of Boundary Scores, see below.)

The results (Fig 5) show that only a tiny proportion of centroids were *not* also Gold Standard names. That is, the vast majority of sets of alternatives in BioCreative do not resemble the (deliberately chosen) AZDC example of Section 3.. Nevertheless, we investigated the reasons why some names were not so well-behaved (see below). We also observe that the recall of Gold Standard names is extremely high. We conclude that any tagger that performs well in finding centroids will perform equally well against the original Gold Standard, as measured by BioCreative's own evaluation.

One unexpected result was that the algorithm uncovered more centroids (18339) than there were core names in BioCreative (18265). How could this be? It turned out that there were entries in the BioCreative Alternative Gene List that overlapped no core name, e.g *PHA* from offsets 175 to 177 in P01842498A0000; *gp0.7* from 30 to 34 in P01310178A0876. Such cases are probably just errors in the BioCreative data[2]. If nothing else, this demonstrates that a Gold Standard in the BioCreative format requires a certain degree of internal consistency checking. By contrast, such an error simply cannot appear in a centroid representation. Every alternative must extend a centroid.

Fig 5 also demonstrates that there are fewer true and false positives according to the evaluation procedure than there are "core" names. This is surprising too. One might expect every candidate name either to be true (match the Gold Standard in some way) or false. This is not so however because BioCreative evaluation actually depends on the distribution of overlapping names between the Gene List and Alternative Gene List. If two candidates match smaller alternatives of the same longer core name, then they count for only *one*, not *two*.

Our qualitative analysis of discrepancies also highlighted this possibility. Fig. 6 shows two such cases. In each case,

[2]Errors because a non-overlapping alternative will not be used by the evaluation script.

db	#cents	TP	FP	FN	P	R	F	#core names
BC	18339	18141	134	124	0.993	0.993	0.993	18265
AZDC	3206	3186	15	32	0.995	0.990	0.993	3218

Figure 5: Centroid evaluation

the core name from the Gene List is shown italicized.

1st Example (superstring is core)
Rel-related human p75 nucleoprotein complex
Rel
Rel-related human p75
Rel-related
human p75 nucleoprotein
human p75 nucleoprotein complex
human p75
p75 nucleoprotein complex
p75
2nd Example (superstring is non-core)
Gamma glutamyl transpeptidase
GGTP
Gamma glutamyl transpeptidase (GGTP)

Figure 6: Disjoint substring alternatives

In the first example, if a candidate annotation includes *Rel* but not *p75*, BioCreative will give no recall penalty. This is because *Rel* and *p75* are just equally good alternatives to annotating the one core entity: *Rel-related human p75 nucleoprotein complex*. If one name in the Alternative Gene List which overlaps it has been found, then the core name has been found. So there is no recall error. By contrast, the centroid algorithm finds that the two smaller names are *both* centroids and evaluation will therefore require both of them to be found by an automatic annotator. This seems to us intuitively correct. Given the alternatives in Fig 6, both *Rel* and *p75* ought to be annotated either individually or as parts of one long name. Otherwise, a use of a gene name has been missed.

Fig 4 shows in xml format the *p75* centroid that is found by the algorithm. It is important to note that the distribution of boundaries does not *tie* particular left boundaries to particular right boundaries. Therefore, *p75 nucleoprotein* is *permissible* given the centroid representation. In this way, the centroid representation further deviates from what is represented in the BioCreative Gold Standard. *p75 nucleoprotein* is not in fact included in the long list of alternative names that are deemed Gold Standard by BioCreative. If an annotation included this candidate, it would simply be false positive. According to the centroid measure, however, it is true positive with a right boundary score (see section 6.) of 0.33.

Who is right? Although no external validation (such as a Mesh identifier) for *p75 nucleoprotein* could be found in a web-search, we note that BioCreative itself deems *human p75 nucleoprotein* to be Gold Standard and it certainly seems that *human* is an optional element. (We could find no identifier for *human p75 nucleoprotein* either.) Therefore, we believe that *p75 nucleoprotein* ought to be acceptable too. It is of course not clear how the lists of alternative names in the BioCreative gold standard were arrived at; but our investigations suggest that they may not be complete.

We carried out the same procedure and analysis for the Arizona Disease Corpus. In this case it was necessary to partition the undifferentiated set of names into 'core' and 'alternative' and this was done without human intervention. Whenever an overlap was detected, the core name was simply the first member of the set. However, the resulting partition only contained 194 alternatives for 3218 core names which is significantly different from BioCreative where 14499 alternatives are associated with 18265 core names. As a result, although the precision, recall and f-scores for AZDC in Figure 5 resemble those for BioCreative, the underlying situation is very different. The centroid algorithm only has work to do when there are alternatives and this is infrequently the case in AZDC.

Qualititative analysis of the AZDC disagreements revealed that they mostly resembled the example of Figure 2. For example, *ovarian cancer* and *cancer cell growth*, are overlapping alternatives in AZDC. This leads to a centroid of *cancer* which is not itself deemed gold standard. In fact, in this particular case, the superstring *ovarian cancer cell growth* was also not AZDC gold standard! One might defend the omission of 'cancer' on its own from the Gold Standard on the grounds that it is too imprecise or vague, even though a disease tagger that spots *cancer* has at least done something better than one that spots nothing at all. It is harder to understand why the superstring *ovarian cancer cell growth* should not be Gold Standard given that both *ovarian cancer* and *cancer cell growth* are Gold Standard.

The conclusion is that, in most cases, where the centroid representation suggests names that differ from those present in the input list of alternatives, the centroid suggested names are correct.

5.1.1. Appositional cases

There is one common case, however, where the centroid algorithm appears, intuitively, to produce incorrect results. These cases involve the linguistic phenomenon of apposition, where two adjacent terms have one referent, such as "C/EBP homologous protein CHOP-10." Here, the following names are Gold Standard alternatives: *CHOP-10*, *protein CHOP-10* and *C/EBP homologous protein*. In these circumstances, the centroid algorithm detects *protein* as a centroid, in addition to *CHOP-10* and *C/EBP homologous*. Since these cases always arise through apposition and involve a highly general term (such as *gene* or *protein*) as head noun, one effective solution is simply to remove centroids that only consist of such a term.

Alternatively, since our final objective is to uncover the entities named and not just their names, information about referents (sometimes called *normalized names*) should be brought to bear. This however represents a future extension of our work.

5.2. Automatic annotation inputs

We also generated centroids for the re-annotations of BioCreative data submitted as part of the CALBC-II [3] challenge competition (Rebholz-Schuhmann et al., 2011). These inputs are not "perfect" of course. The output centroids represent the "hearts" of the competing annotations: a Silver Standard.

5.2.1. Automatic annotation results

Fig 7 shows the (BioCreative-assessed) scores of four different systems supplied by CALBC project partners (p1 to p4) and the centroid scores at three thresholds (t=1 to t=3). For t=1,2 the centroids *outperform* every partner, at least on F-score.

partner	P	R	F
p1	0.710	0.494	0.583
p2	0.690	0.466	0.556
p3	0.761	0.367	0.496
p4	0.888	0.285	0.432
Ce (t=1)	0.611	0.639	0.625
Ce (t=2)	0.772	0.472	0.586
Ce (t=3)	0.865	0.308	0.454

Figure 7: Partner-only harmonization

Fig 8 is for the challenge participants who at least matched a project partner on their BioCreative-assessed score. The result demonstrates that if we extend the participating annotations to all "good" challenge participants, the overall F-Score (64.8) improves significantly (from 62.5) and all but the very top participant are outscored by the centroids.

	P	R	F
p5	0.968	0.676	0.796
p6	0.627	0.609	0.618
p1	0.709	0.494	0.582
p2	0.690	0.463	0.554
p3	0.763	0.371	0.499
p7	0.578	0.374	0.454
p4	0.887	0.282	0.428
Ce (t=2)	0.630	0.667	0.648
Ce (t=3)	0.731	0.576	0.645
Ce (t=4)	0.822	0.467	0.595
Ce (t=5)	0.867	0.344	0.493
Ce (t=6)	0.894	0.205	0.334

Figure 8: Partner-beating CALBC participants

We also experimented with including *all* challenge participants, regardless of their own individual quality. In this case, the quality of the derived centroids no longer increases although, somewhat encouragingly, the F-measure (for thresholds 2 to 5) remains above 50 (data not shown). Furthermore, we also tested the degree of correlation between performance against silver and gold standards. That is, even though by definition the silver standard is not gold, do systems that perform well against a gold standard *also* tend to perform well against a silver standard? For this test, we considered all seventeen submissions, ranked their F-score performance against each of the two standards and calculated a Spearman's rank correlation coefficient. r_s was 0.745 which, if the submissions can be considered random selections from the population of such taggers, is significant at the $p < 0.001$ level (two-tailed test). The correlation is not perfect of course. For example, the system which came top against the gold standard only came fourth against the silver standard whereas the top-ranking system against the silver standard only came third against the gold standard. Nevertheless, the degree of correlation does provide some basis for the hope that a silver standard could stand proxy for a gold standard.

6. Precision, Recall, Boundary Score

Thus far, we have considered centroids mostly as the 'hearts' of annotations. Now, we consider the *distributional* information.

If we replace an existing Gold Standard with a centroid representation, how should we define precision and recall against it? There are several possibilities. We prefer one where a candidate annotation is *true positive* if it is an *extension*[4] of a centroid and that centroid's *maximal extent* is an extension of the annotation. That is, an annotation must include the whole of the heart and neither annotation boundary may lie outside the distribution. Precision is then just the proportion of user annotations that are true positive. Similarly, we define a centroid as *positively found* if there is a candidate annotation which extends it and which does not extend its maximal extent. Recall is the proportion of centroids that are positively found.

Simply put, user annotations should be tested for whether they lie within the *range* allowed by the distribution and whether they extend its *heart*.

We tested empirically whether an F-Score based on these definitions differed from BioCreative F-Score and found only insignificant variation (data not shown).

The superiority of the centroid scoring mechanism lies in two directions. First, it is an intuitively clear measure; whereas the BioCreative measure is not, as shown in section 3. above. Secondly, we can extend the measure to provide a more fine-grained score, a boundary score, for each annotation. A left (right) boundary scores the ratio of its value to the maximum value for any left (right) boundary of the centroid. Boundary scores therefore lie in the range $0 \leq x \leq 1$. In our running example, values of 2 are assigned at right boundary positions +0 and +16. The maximum is 2. So, *TGF-beta* will not only be true positive but score a perfect 1 for its right-boundary, as will *TGF-beta-specific complex*. They are both equally, and perfectly, acceptable. *TGF-beta-specific* is still true positive but scores 0 for its right boundary.

In purely practical terms, we also note that there is considerable advantage to be obtained from an evaluation program which can find and report on *near misses* (there should be an annotation here, but the boundary is perhaps not perfect) as well as *total misses* (even the very heart of the annotation was missed).

[3] `http://www.calbc.eu`

[4] any string x is an extension of itself

7. Comparison to Other Work

There have been several approaches to *flexible* matching against gold standard named entities (see (Nadeau and Sekine, 2009; Tsai et al., 2006) for recent reviews and (Rebholz-Schuhmann et al., 2010) for a later development incorporating the inverse document frequency (idf) scores of the tokens inside named entities). Even in MUC, the earliest of evaluation methodologies, some credit was given to entities that only overlapped a gold standard name so long as the right semantic type (person, organization, etc) was assigned. In some NER competitions, e.g CONLL and IREX, exact matching against a single right answer has remained the standard (Sang and Meulder, 2003; Sekine and Isahara, 2000).

The idea that candidate names must at least contain a gold standard "core term" is one of the options (briefly) canvassed in (Tsai et al., 2006) where it is also correctly noted that many text mining pipelines that deploy NER do not actually require duplication of the sort of human-expert boundary decision-making used in creating the gold standards. (Tsai et al., 2006) however state that "it is only possible to identify core terms by hand" and pay the idea no further attention. We have demonstrated that, given a set of (reasonably complete) alternatives, the core terms can be identified automatically, as can inconsistencies and omissions in the alternatives set. Of course, obtaining the range of alternatives in the first place is not trivial; but it does not require experts to identify the "core" parts of terms. Indeed, it appears advantageous if human expert annotation is envisaged, from the outset, as aiming at gathering justifiable alternatives rather than aiming unrealistically for the single, ideal, expert-agreed boundary. We note also that the centroid scheme include boundaries with *weights*. These could reflect salience amongst experts, as well as salience in a silver standard consensus. The right process for arriving at (good) sets of alternatives appears to be an unexplored research area.

The combination of automatic annotations in order to generate a "silver" standard has been explored before. (Rebholz-Schuhmann et al., 2010) describe a harmonization built from n contributions by a sequence of $n-1$ pairwise alignments using cosine similarity scores over vectors of token idfs. The output, like most gold standards, represents a single choice of 'ideal' term extent and includes no distributional information. In addition, the output of the scheme is order dependent on the sequence of pairs and, since it is also dependent on a particular tokenization and a particular set of idfs, not wholly transparent. Our output is order-independent, not dependent on token idf scores and represents the distribution of terms. We are also, we believe, the first to report qualitative analysis of such a silver standard.

That combinations of results can outperform any individual contributor is of course a very well explored topic in machine learning. In BioCreative, (Smith et al., 2008) were able to train a tagger over a combination of automatic annotation results which outperformed (with statistical significance) the best contributor. Furthermore, they demonstrated evidence that generally weaker systems still added value overall. Although our results mirror these findings to a degree, we note that own work is strictly orthogonal to this line of investigation. Our objective is not to derive the best *tagging system* but to improve the representations within Gold Standards and to develop further the notion of a consensus Silver Standard and evaluate it. Some early investigations into a trained approach towards harmonization appear in (Campos et al., 2011), but the output of the approach once again does not represent distributional information.

8. Conclusion

We have proposed a new representation for names and their alternatives in both Gold and Silver Standard corpora, and an algorithm for deriving them. We have demonstrated that, given gold standard inputs, the outputs perform outstandingly well by traditional measurement whilst also supporting an intuitive picture of the distribution of alternatives and a novel means of assessing the acceptability of variants. Given automatic inputs, the outputs form a *Silver Standard* representing a consensus annotation.

9. Acknowledgments

This work was partially funded by the EU FP7 Support Action grant 231727 (ICT 2007.4.2) *Collaborative Annotation of a Large Biomedical Corpus* and 296410 (ICT 2011.4.1) *Multilingual Annotation of Named Entities and Terminological Resource Acquistion*. We very gratefully acknowledge the data supplied by CALBC challenge II participants which we have used in this study and the insightful comments of our project partners in Erasmus Medical Center, Rotterdam and Friedrich-Schiller University of Jena.

10. References

D. Campos, D. Rebholz-Schuhman, S. Matos, and J.L. Oliveira. 2011. A crf-based approach to harmonize heterogeneous gene/protein annotations. In *Proceedings of the 2nd CALBC workshop*.

C. Kolarik, R. Klinger, C.M. Friedrich, M. Hofmann-Apitius, and J. Fluck. 2008. Chemical names: terminological resources and corpora annotation. *Workshop on building and evaluating resources and corpora annotation, LREC 6, Morocco*.

Martin Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, and Alfonso Valencia. 2008. Evaluation of text-mining systems for biology: overview of the second biocreative community challenge. *Genome Biol.*, 9(suppl 2):S1.

R. Leaman, C. Miller, and G. Gonzalez. 2009. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. *Proc. Pacific Symposium on Biocomputing*, 13:82–89.

David Nadeau and Satoshi Sekine. 2009. A survey of named entity recognition and classification. In Satoshi Sekine and Elisabete Ranchhod, editors, *Named Entities*, pages 3–28. John Benjamins.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno-Yepes, Erik M. van Mulligen, Ning Kang, Jan A. Kors, David Milward, Peter Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, and Udo Hahn. 2010.

The calbc silver standard corpus for biomedical named entities - a study in harmonizing the contributions from four independent named entity taggers. In *LREC-10*.

D. Rebholz-Schuhmann, A. Jimeno-Yepes, C. Li, S. Kafkas, I. Lewin, N. Kang, P. Corbett, D. Milward, E. Buyko, E. Beisswanger, K. Hornbostel, A. Kouznetsov, R. Witte, LaurilaJ.B., B.J.O Baker, C.J Kuo, S. Clematide, G. Rinaldi, R. Farkas, G. Mora, K. Hara, L. Furlong, M. Rautschka, Lara Neves M., A. Pascual-Montano, Q. Wei, N. Collier, F. Mahbub Chowdhury, A. Lavelli, R. Berlanga, R. Morante, V. Van Asch, W. Daelemans, J.L. Marina, E. van Mulligen, J. Kors, and U. Hahn. 2011. Assessment of ner solutions against the first and second calbc silver standard corpus. *Journal of Biomedical Semantics*, 2(Suppl 5:S11).

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147.

Satoshi Sekine and Hitoshi Isahara. 2000. Irex: Ir and ie evaluation project in japanese. In *LREC-2*.

Larry Smith, Lorraine Tanabe, Rie Ando, Cheng J. Kuo, Fang I. Chung, Chun N. Hsu, Yu S. Lin, Roman Klinger, Christoph Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig Struble, Richard Povinelli, Andreas Vlachos, William Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tsai, Hong J. Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel M. Lopez, Jacinto Mata, and John W. Wilbur. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2).

Tzong-Han H. Tsai, Shih-Hung H. Wu, Wen-Chi C. Chou, Yu-Chun C. Lin, Ding He, Jieh Hsiang, Ting-Yi Y. Sung, and Wen-Lian L. Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7(1):92+, February.