Deriving an English Biomedical Silver Standard Corpus for CLEF-ER

Ian Lewin[1] and Simon Clematide[2]

[1] Linguamatics Ltd, 324 Science Park, Milton Road, Cambridge CB4 0WG
ian.lewin@linguamatics.com,
[2] University of Zurich, Binzmühlestr. 14, 8050 Zürich
simon.clematide@uzh.ch

Abstract. We describe the automatic harmonization method used for building the English Silver Standard annotation supplied as a data source for the multilingual CLEF-ER named entity recognition challenge. The use of an automatic Silver Standard is designed to remove the need for a costly and time-consuming expert annotation. The final voting threshold of 3 for the harmonization of 6 different annotations from the project partners kept 45% of all available concept centroids. On average, 19% (SD 14%) of the original annotations are removed. 97.8% of the partner annotations that go into the Silver Standard Corpus have exactly the same boundaries as their harmonized representations.

Keywords: annotation, silver standard, challenge preparation

1 Introduction

CLEF-ER[3], a shared task on multilingual annotation of biomedical named entities is part of the EU ICT project "Multilingual Annotation of Named Entities and Terminology Resources Acquisition" (MANTRA)[4] and is hosted by the "Conference and Labs of the Evaluation Forum"[5] (CLEF).

The CLEF-ER challenge asks participants to find a wide range of biomedical named entities in non-English documents. As a possible (though not absolutely necessary) aid, a set of parallel English language documents is supplied which already contains entity markup. The entities are selected from the Unified Medical Language System (UMLS) [1]. In general, there are more synonyms in UMLS for English than for other languages, and so we expect the English markup to be helpful in finding non-English synonyms not already known to UMLS.

In principle, the challenge can be tackled in a variety of ways. For example, one might search the non-English document for the phrase which best translates the entity markup in the corresponding English document. Alternatively, one might perform entity recognition in the non-English document and then attempt to correlate entities across the pair of multi-lingual documents.

[3] http://www.CLEF-ER.org

[4] http://www.mantra-project.eu

[5] http://www.clef-initiative.eu

The English markup supplied for the CLEF-ER challenge was generated automatically using a Silver Standard methodology in order to harmonize six different annotations from the project partners. This paper explains the process used to generate the Silver Standard and the issues raised during its construction. In the following sections, we first outline the task requirements for the CLEF-ER English silver standard annotation. Next, we discuss how the Silver Standard methodology from the predecessor project "Collaborative Annotation of a Large Scale Biomedical Corpus" (CALBC [2]) was adapted to the new scenario. Finally we present some results and our conclusions.

2 CLEF-ER requirements

The CLEF-ER task scenario imposes a number of requirements which makes the collection of manual and/or Gold Standard annotations (i.e. verified by a reliable procedure incorporating expert opinion) particularly onerous.

1. The concepts to be annotated are
   (a) very large in number
   (b) highly specialist
   (c) highly *diverse* in nature, even though they all fit under the broad heading of 'biomedical'. It is unlikely any one individual will be an expert in all of the included concepts.
2. The document set to be marked up is
   (a) very large in size
   (b) specialist
   (c) highly diverse, ranging from extracts of scientific papers, to drug labels to claims in patent documents.
3. The task requires assignment of concept identifiers (sometimes called "normalization" or "grounding") and not just the identification of names in text, perhaps with an indication of their semantic type.

The concepts are taken from UMLS and are (all) the members of the following specified semantic groups: anatomy, chemicals, drugs, devices, disorders, geographic areas, living beings, objects, phenomena and physiology. The sources of the concepts are MeSH (Medical Subject Headings), MedDRA (Medical Dictionary for Regulatory Activities) and SNOMED-CT (Systemized Nomenclature of Human and Veterinary Medicine). There are 531,466 concepts in total. The concept identifiers to be assigned are UMLS Concept Unique Identifiers (or CUIs).

The document set includes nearly 1.6m sentences (15.7m words) from scientific articles (source: titles of Medline abstracts)[6], 364k sentences (5m words) from drug label documentaion (source: European Medicines Agency)[7] and 120k claims (6m words) from patent documents (source: IFI claims)[8].

[6] http://www.ncbi.nlm.nih.gov

[7] http://www.ema.europa.eu/ema

[8] http://ificlaims.com

```
endogenous TGF-beta-specific complex
           TGF-beta
endogenous TGF-beta
           TGF-beta-specific complex

...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 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 ...
```

Fig. 1. BioCreative alternatives (id:534680) & inter-entity character counts

One additional requirement which clearly distinguishes the CLEF-ER entity recognition task from other similar challenges is that the recognition is to be performed in a different language from that of the supplied marked up data. This has the advantage, for current purposes, that the *precise* boundaries of the entities supplied in the English source data are not of critical importance. No challenge submissions will be evaluated against them. Nevertheless, the assignment of *reasonable* boundaries still remains important to the challenge not least as a possible locus for the use of machine translation technology in finding correlates in other languages. Furthermore, it is important that good precision is obtained in the assignment of CUIs.

These requirements all clearly speak to the use of an automatically derived silver standard for the English annotation. Automatic annotation enables one to cover large amounts of data and, by suitably combining the results of several systems, to avoid precision errors introduced by the inevitable idiosyncracies of any one system. Consequently, we decided to adapt and re-use the centroid harmonization methodology, deployed in a previous large-scale annotation challenge: the CALBC challenge (Collaborative Annotation of a Large Scale Biomedical Corpus).

3 The centroid harmonization of alternative annotations

Figure 1 shows four human expert (BioCreative Gold Standard) annotations over the string *endogenous TGF-beta-specific complex*.

The centroid annotation algorithm, developed initially for the CALBC project, reads in markup from several different annotators and generates a single *harmonized* annotation which represents both the common heart of the set of annotations *and its boundary distribution.* The inputs can be Gold Standard as in Figure 1 or imperfect automatic annotations.

First, text is tokenized at the character level and (ignoring spaces) votes are counted over pairs of *adjacent inter-entity* characters in the mark-up. Figure 1 also shows the inter-entity character counts. For example, only two of the markups consider that the transition *e-n* at the start of *endogenous* falls within an entity. All four consider that *T-G* do. The focus on *inter-entity* pairs, rather than single characters, simply ensures that boundaries are valued when two

Visceral <e b=’l:0:2,l:9:3,r:0:5’> adipose tissue</e>is particularly responsive to somatropin

Fig. 2. Centroid plus boundary distribution markup

different entity names happen to be immediately adjacent to each other in a markup. Over the course of a whole text, the number of votes will mostly be zero, punctuated by occasional bursts of wave-like variation. The *centroids* are the substrings over character pairs that are *peaks* (or local maxima) in a burst of votes. In Figure 1, *TGF-beta* is the centroid.

The inter-entity character votes also define the boundary distribution around the centroid. We define a 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 receives a value of 2 (the difference between 4 and 2) and another before *endogenous*, which also receives 2 (the difference between 0 and 2). Therefore, these alternative boundaries are equally preferred. There is however *no* boundary immediately after *specific*.

With respect to Gold Standard inputs, the advantages of the centroid representation lie first in its perspicuous representation of variability and secondly in its use as something which candidate annotations can be evaluated *against* (see [3] for details) with a clear semantics and scoring method in which equally preferred alternatives are equally scored and more preferred alternatives score more highly than less preferred ones.

When the inputs are less than perfect (the result of automatic annotators rather than human experts), the result is a harmonized *Silver* Standard. Figure 2 shows one centroid uncovered by applying the algorithm to the annotations of CLEF-ER data generated by the five Mantra project partners. In this case, the centroid itself is simply the most highly voted common substring *adipose tissue* but the boundary distribution shows the alternative boundaries.

‘l:0:2” represents the left boundary at exactly the centroid left boundary (position 0) for which two votes were cast. “l:9:3” represents the boundary 9 characters to the left with three votes. “r:0:5” shows that all five votes agreed that the centroid’s right boundary is exactly (position 0) the right boundary of the entity.

One advantage of a centroid Silver Standard is that one easily tailors it for precision or recall (for example, by throwing away centroids or boundaries with very few votes). It should be noted that the centroid heart, *adipose tissue* in no way represents the *correct* annotation, or even the *best* annotation. In the evaluation scheme of [3] it is simply the string which a candidate annotation must at least cover in order to be true-positive; and a candidate which also included *Visceral* would in fact score proportionately more because more standards-contributing annotators agree that it should.

4 Adapting centroids for CLEF-ER

Although we considered giving CLEF-ER participants the maximally informative centroid representation, it was at least equally desirable to provide the *simplest possible* representation. We did not wish to discourage participation. There was also no guarantee that the distributional information could be usefully exploited by participants, nor were we ourselves planning to exploit it in evaluation. Therefore, for clarity and simplicity, we decided to offer individual entity markup in a classical format, rather than a distributional format.

First, we apply a threshold to centroids so that only centroids with at least three votes percolate through to the Silver Standard. Higher thresholds would generate a higher precision harmonization but experiment showed that a three vote threshold gave good recall without sacrificing too much in precision.

Secondly, in the context of CLEF-ER, we decided that, where harmonization offered a distribution over boundaries, it would be more useful to participants to offer the *widest possible boundary*, again subject to a threshold, rather than the most popular boundary. In this way, the greatest amount of lexical content would be included in the marked up entities. Consequently, for each centroid, we calculate an extended centroid, or e-centroid, which has the greatest leftmost (rightmost) boundary with votes above the boundary threshold. The e-centroid for Figure 2 would therefore have a leftmost boundary at position 9, since the leftmost boundary receives the most votes of any left boundary. In this case, the boundary also coincides with the most popular boundary.

Thirdly, we considered the assignment of CUIs to e-centroids. Since e-centroids result from a harmonization of several voting systems, this is not entirely trivial. In the vast majority of cases, the text stretch of the e-centroid does match the text stretch of at least one of the voting systems, in which case those CUIs are assigned. In cases where the string extent of *no* contributing annotation exactly matched an e-centroid, we experimented with a voting system over CUIs. It turned out however that such cases were nearly always the result of unusual combinations of errors from different contributing systems. Therefore, rather than try to assign a CUI, we used the absence of agreement between the e-centroid and *any* voting system as a filter on the silver standard.

4.1 The granularity of harmonization

Finally we re-considered the issue of which annotations to harmonize.

In the CALBC challenge, the task had been to determine typed mentions, i.e. the string extents of entities of certain specified semantic groups. Thus, when evaluating challenge submissions over the title *Bacterial eye infection in neonates, a prospective study in a neonatal unit* looking for `disease` and `anatomy` markup, it makes sense to evaluate against one centroid centred on *infection*, but preferably extending to the left and another centred on *eye* and preferably *not* extending at all. In order to generate these centroids, we run the centroid algorithm once for all annotations of type `disease` and independently for all annotations of type `anatomy`.

<e grp=’ANAT’ cui=’C1563740’>Visceral adipose tissue</e>is particularly <e grp=’DISO’ cui=’C1273518’>responsive to</e><e grp=’CHEM’ cui=’C0376560’>somatropin</e>

Fig. 3. One annotation contributing to silver standard

The Silver Standard CLEF-ER English data is supplied as “hints” on what entities might be found in the non-English text and, since we are not supplying a distribution, it could be misleading to suggest that *eye infection* might be found, when a different disease *bacterial eye infection* can also be found.

The centroid algorithm in itself is perfectly neutral over the range of its inputs, though its outputs will make semantic sense only if the inputs are all annotations *of the same semantic sort.* The *sort* however can be semantic groups, or types or even individual CUIs, with the deciding factor being only a) the intended use of the output centroids b) the possibility of data sparsity in a too fine-grained harmonization. For example, if one harmonizes at the level of CUI and some systems annotate only the longest match and others only the shortest match, there is a clear danger that the minimum thresholds will not be met in *either* case.

We therefore experimented with both types of harmonization. The result was that the gain in the number of entities obtained outweighed the (only small) sparsity issues that resulted, especially as our thresholding was being set to low values (3 votes or more). Therefore, harmonization by CUI was our preferred option and final choice.

5 Results and Examples

To illustrate the sort of work that harmonization carries out, Figure 3 shows the annotations generated by one of the contributing Mantra project partners. One of the other four partners agreed exactly with the annotation over *Visceral adipose tissue*, and another also agreed with the string extent. Consequently, this annotation is propagated to the silver standard, and in fact it would be regardless of whether harmonization were carried out by group or by CUI. Two of the four partners did not support this annotation although they *did* annotate the substring *adipose tissue* and agreed amongst themselves on the CUI for it. Consequently, in harmonization by CUI, this annotation also propagates to the silver standard. In harmonization by group, this annotation would be lost, at least if the “widest boundaries which meet the threshold” is used as the criterion for selecting from the distribution. No other partner agreed with the `DISO` annotation of Figure 3 so this annotation is not replicated in the silver standard. All partners agreed with the string extent of the annotation over *somatropin* but the three other CUI-assigning annotations supported a different CUI. In harmonization by CUI, therefore, this annotation does not propagate, although

Table 1. Absolute number of annotated CUIs provided by the project partners and the percentage of these annotations that are part of harmonized corpora produced with different voting thresholds.

Partner		Voting Threshold						Mean
	Annotations	1	2	3	4	5	6	
P1	6,346,873	99.6%	86.0%	80.0%	72.6%	65.1%	43.8%	78.2%
P2	5,278,663	97.6%	95.9%	90.5%	86.4%	76.7%	52.6%	85.7%
P3	8,595,019	99.6%	86.5%	78.9%	65.8%	55.2%	32.3%	74.0%
P4	5,986,506	99.4%	96.8%	89.6%	77.2%	46.6%	46.5%	79.4%
P5	13,701,552	99.0%	59.7%	51.3%	42.7%	35.8%	20.9%	58.5%
P6	6,482,321	99.5%	97.7%	95.2%	85.5%	72.6%	42.9%	84.8%
Mean	7,731,822	99.1%	87.1%	80.9%	71.7%	58.7%	39.8%	76.8%

another annotation with the same string extent (and with the other CUI) does propagate.

In Table 1 we give a quantitative evaluation of all 6 partner annotations and show how many of these annotations find their way into a harmonized corpus of a given voting threshold. As explained in the last section, there are a few annotations that are lost even at a voting threshold of 1. Partner P2 and P4 annotated relatively few CUIs, however, most of these annotations had a broad support and contributed substantially more CUIs for the harmonized corpora. Table 1 shows also that our final voting threshold of 3 removes on average 19% (SD 14%) of all partner annotations.

A slightly different view on the effects of voting thresholds on our data is presented in Table 2. For each corpus, we evaluate the loss of CUI annotations against a voting threshold of 1. In total, this amounts to 16 million centroids. The stock of centroids having support from all partners (voting threshold 6) is quite low, i.e. the common set of annotations is 17% of the whole set of annotations available from all partners that go into the SSC. For the preparation of the final SSC for the challenge, we inspected in detail the most frequent annotations from the voting thresholds 2 and 3. Raising the threshold from 2 to 3 filtered a substantial amount of noise. Raising the threshold to 4 would have been viable too. However, threshold 3 was more inline with the strategy to stick to a reasonably high recall harmonization.

In Table 3 we investigate the effect of boundary thresholds with respect to the original partner annotations for the final voting threshold of 3. Two questions are addressed here. How many annotations from partners that are covered by a e-centroid retain their original boundaries? How many of them get shortened or extended by different e-centroid boundary thresholds? Between 96.4% and 97.8% of the partner annotations are represented by e-centroids with exactly the same boundaries. For obvious reasons, a boundary threshold of 1 can *only extend* partner annotations. A threshold of 3 leads to a shorter e-centroid representation for the majority of inexact boundary matches. A boundary threshold of 2 represents a balanced strategy and was the setting for the final Silver Standard Corpus.

Table 2. Number of annotated CUIs in the corpora by applying different voting thresholds. There were 6 partner annotations available for the creation of the harmonized corpora.

Voting	EMEA		Medline		Patent		All Corpora	
Threshold	abs	rel	abs	rel	abs	rel	abs	rel
1	2,640,837	100%	11,388,401	100%	2,210,797	100%	16,240,035	100%
2	1,393,936	53%	5,964,424	52%	1,290,346	58%	8,648,706	53%
3	1,147,152	43%	5,043,464	44%	1,087,133	49%	7,277,749	45%
4	963,835	36%	4,140,564	36%	838,000	38%	5,942,399	37%
5	792,271	30%	3,359,090	29%	691,322	31%	4,842,683	30%
6	460,444	17%	1,941,024	17%	375,547	17%	2,777,015	17%

Table 3. Effect of different boundary thresholds (B) on e-centroids (voting threshold 3) compared to the original partner annotations. Label == means exact boundary match on both sides, +- means cases where the partner's left boundary was extended (+) to the left and some material at the end of the partner annotation was shortened (-).

B	==	=+	+=	=-	-=	+-	++	-+	--	Total
1	34,053,526	774,553	475,395			709	27,550	1,036		35,332,769
2	34,791,790	382,792	122,973	121,210	113,015	58,731	3,634	8,795	6,170	35,609,110
3	34,958,525	119,111	103,316	298,907	128,489	105,444	2,876	4,652	7,304	35,728,624

6 Conclusion

The CLEF-ER challenge is an interestingly new variant on the classic "named entity recognition" task. In order to generate good quality English marked up data to be provided as part of the challenge, an automatic annotation method was required. The centroid harmonization method has proved a good basis for building a Silver Standard suitable for the CLEF-ER challenge, even in the circumstance where its distributional nature is not to be exploited in a direct evaluation.

On average, about 81% of all partner annotations are represented by the harmonized annotations using the final voting threshold of 3. This amounts to 45% of all centroids that could have been built by a voting threshold of 1.

For a voting threshold of 3, 97.8% of the partner annotations which go into the Silver Standard Corpus have exactly the same boundaries as their e-centroid representations. A boundary threshold of 2 extends a reasonable amount of the remaining cases with divergent boundaries.

This work is funded by EU Research Grant agreement 296410.

References

1. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32(Database-Issue), 267–270 (2004)

2. D. Rebholz-Schuhmann et al.: Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. Journal of Biomedical Semantics 2 (11/2011 2011), `http://www.jbiomedsem.com/content/2/S5/S11`
3. Lewin, I., Kafkas, S., Rebholz-Schuhmann, D.: Centroids: Gold standards with distributional variation. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul (May 2012)