MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

© 2015 IMIA and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-564-7-643

Identifying Diseases, Drugs, and Symptoms in Twitter

Antonio Jimeno-Yepes[*a,b], Andrew MacKinlay[*a,b], Bo Han[a], Qiang Chen[a]

[a] *Melbourne Research Lab, IBM Research, Victoria, Australia*
[b] *Dept. of Computing and Information Systems, University of Melbourne, Australia*

Abstract

Social media sites, such as Twitter, are a rich source of many kinds of information, including health-related information. Accurate detection of entities such as diseases, drugs, and symptoms could be used for biosurveillance (e.g. monitoring of flu) and identification of adverse drug events. However, a critical assessment of performance of current text mining technology on Twitter has not been done yet in the medical domain. Here, we study the development of a Twitter data set annotated with relevant medical entities which we have publicly released. The manual annotation results show that it is possible to perform high-quality annotation despite of the complexity of medical terminology and the lack of context in a tweet. Furthermore, we have evaluated the capability of state-of-the-art approaches to reproduce the annotations in the data set. The best methods achieve F-scores of 55–66%. The data analysis and the preliminary results provide valuable insights on identifying medical entities in Twitter for various applications.

Keywords:
Social media, data set generation, medical entity recognition.

Introduction

The volume of data on social media sites, such as Twitter, is so vast that it would almost be surprising if it did not contain useful medical information. If we could successfully mine even a small percentage of this, there would be many potential uses, including biosurveillance (e.g. monitoring of seasonal flu) and identification of potential adverse drug events. Previous work exists on biosurveillance based on emergency department notes [1], news data [2], and search data [3], while additional work exists on the detection of adverse effects extracted from forum data [4-6] and Wikipedia [7].

While there has been some work on medical text mining in social media (i.e. identification of relevant tweets for adverse drug events [8]), a critical assessment of performance of current text mining technology has not been performed. It has already been established that Twitter itself presents unique challenges for text mining in the open domain [9,10].

In this work, we have developed an annotated data set from Twitter feeds that can be used to train and evaluate methods to recognise mentions of diseases, symptoms, pharmacologic substances in social media, and particularly microblogs. Furthermore, we have evaluated the performance of existing state-of-the-art entity recognition approaches on this data set. Overall, methods based on conditional random fields allow high precision entity recognition, while additional work is required to improve the recall.

[*] The first and second authors contributed equally to this paper

Materials and Methods

This section presents the development of the Twitter data set and the annotation process, as well as the methods used to automatically reproduce the annotation.

Data Collection And Filtering

We obtained our data using Twitter Stream API from 13/05/2014 to 28/05/2014, collecting 43 million tweets in total. An inspection of random samples indicated that most tweets do not contain medical entities. We pre-filtered tweets using a list of medical terms. We considered three types of entities: *diseases*, *symptoms,* and *pharmacologic substances* to match the particular entities we were targeting for annotation. The list of these biomedical entities comes from the Unified Medical Language System (UMLS) [11] Metathesaurus, which integrates over 100 biomedical terminologies and ontologies. The concepts in the Metathesaurus are assigned one or more semantic types from the UMLS Semantic Network[i] [12]. We downloaded the UMLS version UMLS2014AA and used the default installation. Within this network, we selected the following semantic type mappings:

- T047 (*Diseases or Syndrome*) for diseases,
- T184 (*Sign or Symptom*) for symptoms,
- T121 (*Pharmacologic Substance*) for pharmacologic substances.

From the Metathesaurus, we only retained concepts linked to one of the preceding semantic types, and then extracted the union of terms corresponding to these concepts.[ii] Furthermore, only terms in English and non-obsolete entries were kept. Table 1 shows statistics for each entity type. Disease and pharmacologic substances contain a large number of concepts and terms, while the list of symptoms is comparatively small.

Table 1 – Statistics of the concepts and terms extracted from the UMLS for the three entity types.

Entity	Concept-term pairs	Unique terms	Unique concepts
Disease	201,916	201,013	46,993
Pharm. Subs.	279,261	278,390	120,330
Symptom	19,927	16,865	3,850

[i] UMLS Semantic Network site: http://semanticnetwork.nlm.nih.gov
[ii] We joined the relevant Metathesaurus tables ('MRCONSO' and 'MRSTY') to determine this information.

We filtered the Twitter data to keep the volume manageable. Only tweets with two or more entity types (e.g. a *symptom* and *disease*) were kept, ensuring the filtering is fairly precise (while potentially reducing recall). The lists of UMLS terms extracted, as described above, contain very frequent terms that are primarily used in a non-medical sense, such as *said* and *water*. To avoid a large number of false positives in the filtering stage, we used the frequency of UMLS terms in filtered tweets from the previous step and ranked them in decreasing order of frequency. We then manually removed common terms with a non-medical primary sense from the top 200 terms in each type.

We further removed duplicates in the filtered tweets and removed non-English tweets using LANGID.PY [13]. This filtering pipeline ultimately yielded 11,647 tweets. To investigate whether our filtering (particularly our requirement of the presence of two entity types) erroneously excluded relevant tweets, we examined 1,000 randomly selected tweets. In this sample, we found no relevant tweets—that is, tweets containing biomedical entities of interest—which would have been excluded, suggesting that this pre-filtering methodology is fairly reliable for including all possibly relevant data.

Data Set Annotation Procedure

Four annotators with no medical training annotated the data set with entities from the three semantic types, using BRAT [14]. We created the guidelines iteratively as described here in chronological order.

We prepared an initial set of guidelines based on manual examination of a small subset of the data which was not used for later annotation. We then had all four annotators use these initial guidelines to annotate the same set of 100 tweets as a calibration set.

We checked the inter-annotator agreement on the calibration set and found that it was too low to be appropriate for what we would like to consider a high-quality data set. On the basis of discussions between the annotators, we then refined the guidelines to attempt to resolve frequently observed points of ambiguity in the calibration set. A subsequent round of annotation with the new guidelines improved the inter-annotator agreement, but still not to a level we considered acceptable. At this point we moved to a system of having all tweets annotated by two annotators, then merging the annotations and having the annotators resolve all disagreements in discussion.

This double annotation methodology led to a higher-quality data set with very respectable figures for inter-annotator agreement as shown below. The cost, of course, was that we were able to obtain less data per hour of annotation time. In practice we found the merging and discussion process was generally fast (roughly 20% of the annotation time itself), meaning that overall annotation efficiency was reduced by a factor of around 2.4 by the double annotation methodology. However, for future iterations of the data set, we will leave open the option of augmenting this high quality data set with a lower-quality data set annotated only by a single annotator (possibly assisted by automatic pre-annotation).

Since we settled on double-annotation, the relevant comparison for inter-annotator agreement is to calculate agreement on a subset of the data, which has been annotated twice in the manner previously described. That is, the four annotators were grouped into two pairs, and each pair annotated and merged the same subset of 100 tweets using the methodology described above. We then calculated agreement figures between the two sets of annotations, obtained using BratEval[iii] [15].

Table 2 – Inter annotator agreement for each one of the entity types.

Entity	Precision	Recall	F1
Disease	0.8400	0.8750	0.8571
Pharm. Subs.	0.9500	0.8261	0.8837
Symptom	0.8246	0.8393	0.8319

Inter annotator agreement, as shown in Table 2, is reassuringly high, particularly for such a potentially ambiguous task, and displayed similar levels of agreement to other biomedical annotation tasks [16]. Most of the disagreements were terms inadvertently missed by the annotators, and in a few cases the words were arguably non-medical terms, which can sometimes be difficult to distinguish. For instance, terms like *chill* or *weak* often carry a non-medical meaning. Other disagreements included different interpretations of the subtleties of the guidelines, such as whether *pill* was sufficiently specific to be included.

Annotation Guidelines

After the two rounds of annotation calibration, we settled on a final set of annotation guidelines. These stipulated that we are interested in annotating three kinds of entities: pharmacologic substances, diseases, and symptoms. In addition to traditional entities, which correspond to noun phrases, we also broadened the scope of the annotation to allow for short phrases headed by verbs (such as *I **coughed** all morning*) and adjectives (such as *felt **light-headed***), which indicate diseases or symptoms. If the part-of-speech of the head-word of an annotated item is not a noun, the annotation was marked as an adjective or verb as appropriate in a separate attribute.

We also found that in many cases a concept being mentioned, which may have looked superficially medical, was unlikely to truly refer to a medical concept. In particular, mentions may be metaphorical, figurative, or purely humorous, and in these cases annotators were instructed to apply the 'figurative' attribute. There was also another slightly distinct class of items, one in which the terms have an informal and non-clinical meaning in addition to the clinical one. If the non-clinical meaning is clearly being used, annotators were to apply the 'non-medical' attribute, such as in ***depressed** about my exam results*.

The guidelines also instruct annotators to annotate the most specific entity possible (e.g. *codeine syrup* rather than *syrup*) and to include as many tokens as possible as long as they are part of a fixed expression referring to a particular kind of entity (e.g. *disgusting* would not be part of the entity in *disgusting codeine syrup*). It was also specified that entities which do not distinguish anything more specific than the base entity category should not be annotated, as the very general information in these is unlikely to be specific enough to be useful in downstream applications. In addition, it was also permitted to have overlapping annotations, so while *pain medication* would most sensibly be annotated as a pharmacologic substance, the token *pain* within it should also be annotated as a symptom.

Correctly identifying medical concept mentions and categorising them is sometimes a difficult task for annotators without formal medical training as there is a large terminology space. So it may be difficult to determine whether a particular token refers to a valid concept (e.g. Should *prune juice* be

[iii] BratEval site: https://bitbucket.org/nicta_biomed/brateval

considered a pharmacologic substance? Is *dextromethorphan* a real drug name?), or which of two categories it should refer to. In particular, the division between the disease and symptom categories can be very uncertain in many cases. For example, it may not be clear initially whether *amnesia* is a disease or a symptom. So, the annotators were advised to refer to the UMLS in cases of uncertainty, essentially using the UMLS as a substitute for in-depth domain knowledge. In particular, the UMLS semantic type is important, so generally the semantic types should obey the same mapping as described in 'Data collection and filtering' (for example, *disease* entities should have semantic type 'T047').

Even if the UMLS was not the perfect resource for making these decisions, it is in widespread usage (incorporating many standardised terminologies), and at least provides a common basis for decision-making, ensuring consistency of the annotations. In some cases, the context may make it clear that a strict interpretation of UMLS semantic type, as described above, is not appropriate; in these cases, annotators were free to apply their own more appropriate categorisation instead.

It is of course difficult to codify every possible annotation decision in a static set of guidelines. Inevitably, certain borderline cases had to be decided on the basis of the annotators' intuitions in ways that are difficult to encode specifically. However, the procedure of double annotation that we adopted and the high inter-annotator agreement we achieved at least suggests that the annotations are reasonably internally consistent, and thus presumably repeatable.

Data Set Statistics

The final data set contains 1,300 annotated tweets in 13 files with 100 tweets each. As shown in the following table, *symptom* is the most frequent type with no significant difference between disease and pharmacologic substance. We can see that entities are typically composed of a single token for *symptoms*, while *diseases* and *pharmacologic substances* more frequently span multiple tokens.

Table 3 – Statistics for the entities in the data set. Entities annotated as non-medical have not been considered.

Entity	No Entities	Avg. length	Avg. Tokens
Disease	253	10.40 ± 5.83	1.41 ± 0.62
Pharm. Subs.	233	9.83 ± 4.35	1.39 ± 0.58
Symptom	764	6.66 ± 2.96	1.13 ± 0.40

Table 4 shows the number of non-medical and figurative terms annotated. The number of pharmacologic substances and symptoms that are not medically related are significantly larger; it seems these terms are more often used informally, rather than with their medical definition. In addition, almost 10% of the symptom mentions are used figuratively.

Table 4 – Number of non-medical and figurative entities.

Entity	Non-medical	Figurative	Medical Figurative
Disease	1 (0.40%)	5 (1.98%)	4 (1.58%)
Pharm. Subs.	20 (8.58%)	2 (0.86%)	2 (0.86%)
Symptom	122 (9.65%)	124 (9.81%)	44 (3.48%)

Table 5 – Part-of-speech of the annotated entities. Entities annotated as non-medical have not been considered.

Entity	Noun	Adjective	Verb
Disease	246 (97.2%)	7 (2.8%)	0 (0.0%)
Pharm. Subs.	233 (100.0%)	0 (0.0%)	0 (0.0%)
Symptom	454 (75.5%)	262 (20.7%)	48 (3.8%)

Almost all entities annotated are nouns as noted in Table 5. It is logical that pharmacologic substances are nouns. There are only a few mentions of diseases that appear as adjectives (e.g. *blind*, *obese,* or *overweight*). For symptoms, over 20% appear as adjectives (e.g. *breathless*, *hungry,* or *sick*) and a smaller quantity appear as verbs (e.g. *fainted*, *coughing,* or *shaking*).

Table 6 shows the top 10 terms by frequency per entity type. There is a large number of individual posting on Twitter and we can identify multiple topics in our data set, which shows the possibilities of exploiting Twitter data and the complexity of extracting relevant signals from it. Some of these terms denote concerns about diseases that affect a large part of the population (e.g. *diabetes*) but also highlight recent breakthroughs in medicine (e.g. a new *malaria vaccine*). We also find mentions of recreational drugs (such as *marijuana* and *cannabis*), which are not related to any specific news item. In addition, there are mentions like *heroin* linked to news when the tweet was posted (e.g. *NYPD officers to carry heroin overdose antidotes*). Furthermore, we find a large number of symptoms that are not linked to any specific disease, which could be monitored as signals for biosurveillance.

Table 6 – Top most frequent terms per entity type.

Disease		Phar. Sub.		Symptom	
diabetes	14	marijuana	12	tired	136
heart disease	10	cannabis	12	pain	93
stroke	10	alcohol	11	hungry	61
cold	8	heroin	8	stress	50
asthma	6	pain meds	7	headache	36
malaria	6	vitamin c	4	sore	15
allergy	4	chill pill	4	sick	14
migraine	4	malaria vaccine	4	cough	13
aids	4	caffeine	4	exhausted	12
obesity	4	calcium	4	hangover	11

Named Entity Recognition

After annotating a data set, we investigated how applicable standard approaches to named entity recognition (NER) would be to this data. This indicates how unique the data is, and helps us to predict how difficult it will be to reliably reproduce these annotations automatically on unseen data, which is our ultimate goal. Three methods were used to annotate the data set with the three entity types.

The first uses MetaMap [16], which is developed at the US National Library of Medicine that maps spans of text to UMLS Metathesaurus concepts and is considered state-of-the-art for this task. It uses parsing to identify spans of text in which entities could appear and then smart dictionary matching to identify the concepts [16]. We used MetaMap 2013, with its default configuration, to perform experiments with and without word sense disambiguation (WSD) [18]. MetaMap output was filtered to keep only concepts belonging to the three semantic types under consideration, and the output was converted to BRAT standoff format for evaluation.

In addition to MetaMap, we considered two systems based on machine learning (ML). Firstly, we created a custom NER tagger for this data set, called 'Micromed' (since it tags medical concepts in microblog posts), to provide an in-house solution. It uses conditional random fields (CRF) [19] as its underlying machine learning algorithm. CRFs are frequently used in state-of-the-art named entity recognizers, including those in the biomedical domain. For its CRF implementation, Micromed uses CRFSuite [20]. The features were derived

*Table 7 – Results of all automatic annotators over all entity types. Statistical significance at p < 0.01: * vs MetaMap; † vs SNER.*

		Disease			Pharm. Substance			Symptom		
	Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Exact Match	MetaMap +WSD	0.4123	0.5850	0.4837	0.2264	0.5150	0.3145	0.5337	0.5604	0.5467
	MetaMap	0.3437	0.7876	0.4786	0.2225	0.7785	0.3460	0.4644	0.7635	0.5775
	Stanford NER	0.7917	0.3071	0.4312	0.8952	0.3565	0.4946*	0.7526	0.5763	0.6509*
	Micromed	0.7987	0.5020	0.6165*†	0.8142	0.3948	0.5318*	0.7193	0.6028	0.6559*
	Micromed +Meta	0.8049	0.5217	**0.6331*†**	0.8205	0.4120	**0.5486***	0.7220	0.6041	**0.6578***
Partial Match	MetaMap +WSD	0.4457	0.6299	0.5220	0.2642	0.5957	0.3660	0.4424	0.6150	0.5146
	MetaMap	0.3437	0.7876	0.4786	0.2225	0.7785	0.3460	0.4644	0.7603	0.5766
	Stanford NER	0.7917	0.3071	0.4312	0.8952	0.3565	0.4946*	0.7603	0.6013	0.6696*
	Micromed	0.7987	0.5020	0.6165*†	0.8142	0.3948	0.5318*	0.7439	0.6234	0.6783*
	Micromed +Meta	0.8049	0.5217	**0.6331*†**	0.7220	0.6041	**0.6578***	0.7450	0.6234	**0.6788**

from those commonly used in NER, and tuned somewhat to this particular task:

- Part-of-speech tag and relative position in a context window of three tokens each side
- Token surface form and relative position, in a context window of two tokens each side
- Token prefix and suffix character N-grams of all lengths up to eight
- Whether the token appears in a list of synonyms for concepts with the appropriate semantic type extracted from UMLS
- (In some configurations) whether MetaMap annotates the token as a concept with that semantic type

Since the goal was to evaluate how similar the task was to standard NER, we did not add particularly radical features. For tokenising and POS-tagging, we used TweetNLP [21]. We trained a CRF model for each category (*diseases*, *pharmacologic substances,* and *symptoms*), treating them as distinct annotation tasks. The output of the CRF engine was converted to BRAT standoff[iv] format for evaluation.

Secondly, we used the Stanford NER tagger (SNER) [22] as another strong baseline to study how difficult the task is for existing NER tools and help identify effective features. This also underlyingly uses a CRF. We reformatted annotations into SNER format and applied the limited default features to train taggers for each of the three categories, which included character n–grams and word tokens in fixed context windows (the primary feature difference from Micromed being the absence of custom lookup lexicons).

Results

To compare the performance of the annotation methods, we used two different evaluation profiles. Exact match requires a given entity from the classifier output to have the same start and end span as the reference entity to be considered a match. Partial match considers entities as matching if there is any overlap at all between the entity produced by the classifier and the gold standard. After counting matches, we calculated precision, recall, and F1 in the usual way. Entity annotations marked as non-medical were ignored in training and testing. Statistical significance of the results shown in Table 7 was computed using a two-sample *t*-test with randomization over the cross-validation folds.

Machine-learning methods were trained and evaluated using 13-fold cross validation, based on 13 sections of the data set, each containing 100 tweets. That is, at each of the 13 iterations, 1200 tweets (12 sections) were used for training, while the remaining 100 tweets were used for evaluation. MetaMap, however, was simply applied to the whole data set, so the results are comparable. We evaluated Micromed, both with and without features based on MetaMap (the former case is denoted "+Meta"), to evaluate how well it could perform without relying on an external tool with a significant overhead. Results of the overall methods are presented in Table 7.

MetaMap results without WSD have much higher recall, but poorer precision. Entities missed include terms that are not in the UMLS (e.g. *painkillers addiction*) or terms in the UMLS that are not in the categories of interest (e.g. *cold* as *symptom*). False positives include non-specific terms annotated by MetaMap (e.g. *drugs*), which are excluded by the annotation guidelines, terms bearing a non-medical meaning (e.g. *I'm sick and tired of negativity),* and WSD mistakes (e,g. *cannabis* was annotated as *plant* instead of *substance)*.

The ML methods usually outperforms MetaMap. SNER's accuracy is lower over *Disease* but substantially higher over *Pharmacologic Substances* and *Symptom*. Micromed has higher F-score again, with statistically significant increases over SNER except for *Symptom.* SNER has lower recall since it lacks the implicit domain knowledge from the UMLS-derived features of Micromed and Micromed+Meta (including MetaMap). We know, from Table 1, that the *Symptom* category has a smaller vocabulary; thus, relevant information can be learned from the training data alone. Moreover, some symptom entities, which SNER detects, are missed by Micromed; in many cases, these are not in the UMLS (e.g. *magnesium-deficient*). The increase in recall of partial match compared to exact match is not especially significant, except for *Symptom*.

Discussion

The data set was annotated with high inter-annotator agreement. Extra considerations were required for Twitter compared to biomedical literature, e.g., the *Figurative* attribute and extending the concept of entities beyond nouns, which might make traditional NER approaches perform poorly.

MetaMap was not trained on this data set as the other two methods were, but still shows a competitive performance with higher recall, while Micromed and SNER are generally more precision-biased and have a higher F-score overall. The

[iv] Brat standoff format: http://brat.nlplab.org/standoff.html

difference in performance of the methods is scientifically interesting in what it tells us about the nature of the data set. MetaMap is widely used for medical NER due to the respectable performance it achieves over research articles and clinical text. MetaMap has not been tuned for Twitter data and was outperformed in this work by ML-based classifiers, which could be effectively trained on a relatively small in-domain corpus. SNER even lacked in-domain knowledge, while Micromed used mostly standard NER features. This difference between MetaMap and ML methods suggests that the data set here has different characteristics to NER tagging in other domains.

The difference may also reflect different design considerations to be considered to tune MetaMap to work with Twitter data. In our work, the current performance of Micromed, as well as its lower computational cost[v] are likely to be advantageous in processing large volumes of social media data. The computational cost must, of course, be low if we hope to process a significant fraction of a high-volume data stream in real time. In addition, the smaller number of false positives generated by a higher precision system should lead to greater acceptance by users of the system output. Since the volume of data available in social media is relatively large with a high degree of redundancy with respect to overall trends, slightly lower recall is less of a concern in these applications.

Conclusion

In this work, we have presented the development of a Twitter set annotated with medical entities, showing that it is possible to perform high-quality annotation despite the complexity of medical terminology and the lack of context in a tweet. We have made the dataset publicly available[vi] to encourage further research. Furthermore, we have evaluated the capability of some state-of-the-art approaches to reproduce the manual annotations. The approaches demonstrate reasonable accuracy (with interesting variations between the methods), although further work is needed to identify additional features that might improve the performance of the annotators. We have focused on creating a data set and evaluating state-of-the-art annotators. We plan to use these annotators to process a live data stream from Twitter or some other source for biosurveillance and detecting adverse drug reactions.

References

[1] Espino, JU, Wagner MM, Tsui FC, et al. (2004). The RODS Open Source Project: removing a barrier to syndromic surveillance. Medinfo 2004: 1192-6.

[2] Collier N, et al. BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics 2008: 24(24): 2940-2941.

[3] Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. Nature 2008: 457(7232): 1012-1014.

[4] Segura-Bedmar I, de la Pena S, and Martínez P. Extracting drug indications and adverse drug reactions from Spanish health social media. ACL 2014, 98

[5] Metke-Jimenez A, Karimi S, and Paris C. Evaluation of text-processing algorithms for adverse drug event extraction from social media. Proceedings of the first international workshop on Social media retrieval and analysis. ACM, 2014.

[6] Cameron D, et al. "PREDOSE: A semantic web platform for drug abuse epidemiology using social media." Journal of Biomedical Informatics 2013: 46(6): 985-997.

[7] Generous N, et al. "Global disease monitoring and forecasting with Wikipedia." PLoS Computational Biology 2014: 10(11): e1003892.

[8] Ginn R, Pimpalkhute P, Nikfarjam A, et al.. Mining Twitter for Adverse Drug Reaction Mentions: a Corpus and Classification Benchmark, BioTxtM, LREC, 2014

[9] Baldwin T, Cook P, Lui M, et al. (2013). How noisy social media text, how diffrnt social media sources. In Proceedings of IJCNLP 2013.

[10] Han B, and Baldwin T (2011, June). Lexical normalisation of short text messages: makn sens a# twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 368-378)

[11] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Research 2004: 32 (Suppl 1): D267-D270.

[12] McCray, AT. An upper-level ontology for the biomedical domain. Comparative and Functional Genomics 2003: 4(1): 80-84.

[13] Lui M, and Baldwin T. langid. py: An off-the-shelf language identification tool. Proceedings of the ACL 2012 System Demonstrations.

[14] Stenetorp P, Pyysalo S, Topić G, et al. brat: a Web-based Tool for NLP-Assisted Text Annotation. in *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, 2012.

[15] Verspoor K, Jimeno-Yepes AJ, Cavedon L., et al. (2013). Annotating the biomedical literature for the human variome. Database 2013: bat019.

[16] Doğan, RI, Leaman R, and Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of Biomedical Informatics 2014: 47: 1-10.

[17] Aronson AR, and Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 2010: 17(3), 229-236.

[18] Jimeno-Yepes A, and Aronson AR. Integration of UMLS and MEDLINE in unsupervised word sense disambiguation. AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text. 2012.

[19] Lafferty J, McCallum A, and Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. (2001).

[20] Naoaki O. CRFsuite: a fast implementation of conditional random fields (CRFs). URL http://www. chokkan. org/software/crfsuite (2007).

[21] Owoputi O, O'Connor B, Dyer C, Gimpel K, Schneider N, and Smith NA. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In Proceedings of NAACL 2013.

[22] Finkel JR, Grenager T, and Manning C. Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of ACL 2005.

Address for correspondence

Antonio Jimeno Yepes, IBM Research, Lvl 5, 204 Lygon street, Carlton, VIC, 3053, Australia – antonio.jimeno@au1.ibm.com

[v] Single-threaded processing time with MetaMap for this entire data set was 396 CPU-seconds, while Micromed over the same data took 19 CPU-seconds. However, it should be noted that MetaMap provides richer information than Micromed.

[vi] At https://github.com/IBMMRL/medinfo2015