Annotating named entities in clinical text by combining pre-annotation and active learning

Maria Skeppstedt
Dept. of Computer and Systems Sciences (DSV)
Stockholm University, Forum 100, 164 40 Kista, Sweden
mariask@dsv.su.se

Abstract

For expanding a corpus of clinical text, annotated for named entities, a method that combines pre-tagging with a version of active learning is proposed. In order to facilitate annotation and to avoid bias, two alternative automatic pre-taggings are presented to the annotator, without revealing which of them is given a higher confidence by the pre-tagging system. The task of the annotator is to select the correct version among these two alternatives. To minimise the instances in which none of the presented pre-taggings is correct, the texts presented to the annotator are actively selected from a pool of unlabelled text, with the selection criterion that one of the presented pre-taggings should have a high probability of being correct, while still being useful for improving the result of an automatic classifier.

1 Introduction

One of the key challenges for many NLP applications is to create the annotated corpus needed for development and evaluation of the application. Such a corpus is typically created through manual annotation, which is a time-consuming task. Therefore, there is a need to explore methods for simplifying the annotation task and for reducing the amount of data that must be annotated.

Annotation can be simplified by automatic pre-annotation, in which the task of the annotator is to improve or correct annotations provided by an existing system. The amount of data needed to be annotated can be reduced by active learning, i.e. by actively selecting data to annotate that is useful to a machine learning system. When using pre-tagged data, the annotator might, however, be biased to choose the annotation provided by the pre-tagger. Also, if the produced pre-taggings are not good enough, it is still a time-consuming task to correct them or select the correct tagging among many suggestions.

Consequently, there is a need to further explore how an annotated corpus can be expanded with less effort and using methods that will not bias the annotators.

2 Background

The background discusses basic ideas of pre-annotation and active learning, as well as the particular challenges associated with annotating clinical text.

2.1 Annotating clinical text

A number of text annotation projects have been carried out in the clinical domain, some of them including annotations of clinical named entities, such as mentions of symptoms, diseases and medication. Such studies have for example been described by Ogren et al. (2008), Chapman et al. (2008), Roberts et al. (2009), Wang (2009), Uzuner et al. (2010), Koeling et al. (2011) and Albright et al. (2013).

As in many specialised domains, expert annotators are typically required to create a reliable annotated clinical corpus. These expert annotators are often more expensive than annotators without the required specialised knowledge. It is also difficult to use crowdsourcing approaches, such as using e.g. Amazon's Mechanical Turk to hire online annotators with the required knowledge (Xia and Yetisgen-Yildiz, 2012). A further challenge is posed by the content of the clinical data, which is often sensitive and should therefore only be accessed by a limited number of people. Research community annotation is consequently another option that is not always open to annotation projects in the clinical domain, even if there are examples of such community annotations also for clinical text, e.g. described by Uzuner et al. (2010).

Proceedings of the ACL Student Research Workshop, pages 74–80,
Sofia, Bulgaria, August 4-9 2013.

To simplify the annotation process, and to minimise the amount of annotated data is therefore even more important for annotations in the clinical domain than for annotation in general.

2.2 Pre-annotation

A way to simplify annotation is automatic pre-annotation (or pre-tagging), in which a text is automatically annotated by an existing system, before it is given to the annotator. Instead of annotating unlabelled data, the annotator either corrects mistakes made by this existing system (Chou et al., 2006), or chooses between different taggings provided by the system (Brants and Plaehn, 2000). The system providing the pre-annotations could be rule- or terminology based, not requiring annotated data (Mykowiecka and Marciniak, 2011), as well as a machine learning/hybrid system that uses the annotations provided by the annotator to constantly improve the pre-annotation (Tomanek et al., 2012). There exist several annotation tools that facilitate the use of pre-annotation by allowing the user to import pre-annotations or by providing pre-annotation included in the tools (Neves and Leser, 2012).

A condition for pre-annotation to be useful is that the produced annotations are good enough, or the effect can be the opposite, slowing the annotators down (Ogren et al., 2008). Another potential problem with pre-annotation is that it might bias towards the annotations given by the pre-tagging, for instance if a good pre-tagger reduces the attention of the annotators (Fort and Sagot, 2010).

2.3 Active learning

Active learning can be used to reduce the amount of annotated data needed to successfully train a machine learning model. Instead of randomly selecting annotation data, instances in the data that are highly informative, and thereby also highly useful for the machine learning system, are then actively selected. (Olsson, 2008, p. 27).

There are several methods for selecting the most informative instances among the unlabelled ones in the available pool of data. A frequently used method is uncertainty sampling, in which instances that the machine learner is least certain how to classify are selected for annotation. For a model learning to classify into two classes, instances, for which the classifier has no clear preference for one of the two alternatives, are chosen for annotation. If there are more than two classes, the confidence for the most probable class can be used as the measure of uncertainty. Only using the certainty level for the most probable classification means that not all available information is used, i.e. the information of the certainty levels for the less probable classes. (Settles, 2009)

An alternative for a multi-class classifier is therefore to instead use the difference of the certainty levels for the two most probable classes. If c_{p1} is the most probable class and c_{p2} is the second most probable class for the observation $\boldsymbol{x_n}$, the margin used for measuring uncertainty for that instance is:

$$M_n = P(c_{p1}|\boldsymbol{x_n}) - P(c_{p2}|\boldsymbol{x_n}) \qquad (1)$$

An instance with a large margin is easy to classify because the classifier is much more certain of the most probable classification than on the second most probable. Instances with a small margin, on the other hand, are difficult to classify, and therefore instances with a small margin are selected for annotation (Schein and Ungar, 2007). A common alternative is to use entropy as an uncertainty measure, which takes the certainty levels of all possible classes into account (Settles, 2009).

There are also a number of other possible methods for selecting informative instances for annotation, for instance to use a committee of learners and select the instances for which the committee disagrees the most, or to search for annotation instances that would result in the largest expected change to the current model (Settles, 2009).

There are also methods to ensure that the selected data correctly reflects the distribution in the pool of unlabelled data, avoiding a selection of outliers that would not lead to a correct model of the available data. Such methods for structured prediction have been described by Symons et al. (2006) and Settles and Craven (2008).

Many different machine learning methods have been used together with active learning for solving various NLP tasks. Support vector machines have been used for text classification (Tong and Koller, 2002), using properties of the support vector machine algorithm for determining what unlabelled data to select for classification. For structured output tasks, such as named entity recognition, hidden markov models have been used by Scheffer et al. (2001) and conditional random fields (CRF) by Settles and Craven (2008) and Symons et al. (2006).

Olsson (2008) suggests combining active learning and pre-annotation for a named entity recognition task, that is providing the annotator with pre-tagged data from an actively learned named entity recogniser. It is proposed not to indiscriminately pre-tagg the data, but to only provide those pre-annotated labels to the human annotator, for which the pre-tagger is relatively certain.

3 Method

Previous research on pre-annotation shows two seemingly incompatible desirable properties in a pre-annotation system. A pre-annotation that is not good enough might slow the human annotator down, whereas a good pre-annotation might make the annotator lose concentration, trusting the pre-annotation too much, resulting in a biased annotation. One possibility suggested in previous research, is to only provide pre-annotations for which the pre-annotation system is certain of its classification. For annotations of named entities in text, this would mean to only provide pre-tagged entities for which the pre-annotations system is certain. Such a high precision pre-tagger might, however, also bias the human annotator towards not correcting the pre-annotation.

Even more incompatible seems a combination between pre-annotation and active learning, that is to provide the human annotator with pre-tagged data that has been selected for active learning. The data selected for annotation when using active learning, is the data for which the pre-annotator is most uncertain and therefore the data which would be least suitable for pre-annotation.

The method proposed here aims at finding a way of combining pre-annotation and active learning while reducing the risk of annotation bias. Thereby decreasing the amount of data that needs to be annotated as well as facilitating the annotation, without introducing bias. A previous version of this idea has been outlined by Skeppstedt and Dalianis (2012).

The method is focused on the annotation of named entities in clinical text, that is marking of spans of text as well as classification of the spans into an entity class.

3.1 Pre-annotation

As in standard pre-annotation, the annotator will be presented with pre-tagged data, and does not have to annotate the data from scratch.

To reduce the bias problem that might be associated with pre-tagging, the mode of presentation will, however, be slightly different in the method proposed here. Instead of presenting the best tagging for the human annotator to correct, or to present the n best taggings, the two best taggings produced by a pre-tagger will be presented, without informing the annotator which of them that the pre-tagger considers most likely.

When being presented with two possible annotations of the same text without knowing which of them that the pre-annotation system considers as most likely, the annotator always has to make an active choice of which annotation to choose. This reduces the bias to one particular pre-annotation, thereby eliminating a drawback associated with standard pre-annotation. Having to consider two alternatives might add cognitive load to the annotator compared to correcting one alternative, but ought to be easier than annotating a text that is not pre-tagged.

The reason for presenting two annotations, as opposed to three or more, is that it is relatively easy to compare two texts, letting your eyes wander from one text to the other, when you have one comparison to make. Having three optional annotations would result in three comparisons, and having four would result in six comparisons, and so on. Therefore, having two optional annotations to choose from, reduces the bias problem while at the same time still offering a method for speeding up the annotation.

A simple Java program for choosing between two alternative pre-annotated sentences has been created (Figure 1). The program randomly chooses in which of the two text boxes to place which pre-annotation. The user can either choose the left or the right annotation, or that none of them is correct.

The data will be split into sentences, and one sentence at time will be presented to the annotator for annotation.

3.2 Active learning

To choose from two presented annotations might also potentially be faster than making corrections to one presented annotation. For this to be the case, however, one of the presented annotations has to be a correct annotation. In order to achieve that, the proposed method is to use a version of active learning.

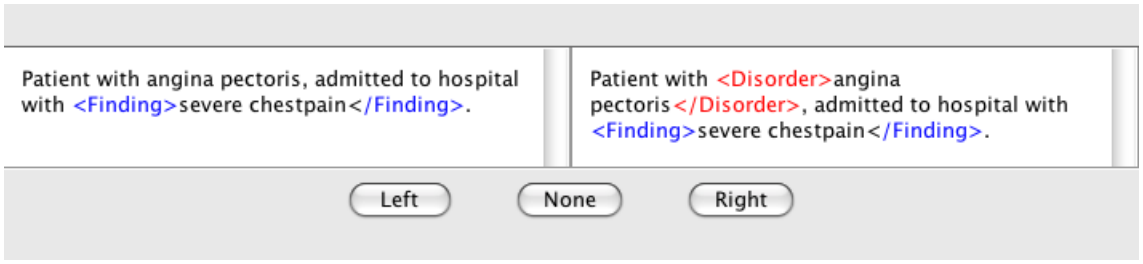

Figure 1: A simple program for choosing between two alternative annotations, showing a constructed example in English.

The standard use of active learning is to actively select instances to annotate that are useful to a machine learner. Instances for which the machine learning model can make a confident classification are not presented to the annotator, as these instances will be of little benefit for improving the machine learning system.

The version of active learning proposed here is retaining this general idea of active learning, but is also adding an additional constraint to what instances that are actively selected for annotation. This constraint is to only select text passages for which it is probable that one of the two best pre-taggings is correct, i.e. the pre-tagger has to be confident that one of the two presented pre-annotations is correct, but it should be uncertain as to which one of them is correct.

For ensuring that the sentences selected for annotation are informative enough, the previously described difference of the certainty level of the two most probable classes will be used. The same standard for expressing margin as used in (1), can be used here, except that in (1), c_{p1} and c_{p2} stand for classification of one instance, whereas in this case the output is a sequence of labels, labelling each token in a sentence. Therefore, c_{p1} and c_{p2} stand for the classification of a sequence of labels.

Let c_{p1} be the most probable labelling sequence, c_{p2} the second most probable labelling sequence and c_{p3} the third most probable labelling sequence. Moreover, let $\boldsymbol{x_n}$ be the observations in sentence n, then the following margins can be defined for that sentence:

$$M_{toSecond\_n} = P(c_{p1}|\boldsymbol{x_n}) - P(c_{p2}|\boldsymbol{x_n}) \quad (2)$$

$$M_{toThird\_n} = P(c_{p1}|\boldsymbol{x_n}) - P(c_{p3}|\boldsymbol{x_n}) \quad (3)$$

To make the probability high that one of the two presented pre-annotations is correct, the same method that is used for determining that an annotation instance is informative enough could be used. However, instead of minimising the margin between two classification instances, it is ensured that the margin in high enough. That is, the difference in certainty level between the two most probable annotations and the third most probable must be high enough to make it probable that one of the two best classification candidates is correct. This can be achieved by forcing $M_{toThird}$ to be above a threshold, t.

The criteria for selecting the next candidate sentence to annotate can then be described as:

$$\boldsymbol{x}^* = \underset{x}{\operatorname{argmin}}\, P(c_{p1}|\boldsymbol{x}) - P(c_{p2}|\boldsymbol{x}) \quad (4)$$

where

$$P(c_{p1}|\boldsymbol{x}) - P(c_{p3}|\boldsymbol{x}) > t$$

As instances with the highest possible $P(c_{p2}|\boldsymbol{x})$ in relation to $P(c_{p1}|\boldsymbol{x})$ are favoured, no threshold for the margin between $P(c_{p2}|\boldsymbol{x})$ and $P(c_{p3}|\boldsymbol{x})$ is needed.

It might be difficult to automatically determine an appropriate value of the threshold t. Therefore, the proposed method for finding a good threshold, is to adapt it to the behaviour of the annotator. If the annotator often rejects the two presented pre-taggings, text passages for which the pre-tagger is more certain ought to be selected, that is the value of t ought to be increased. On the other hand, if one of the presented pre-taggings often is selected by the annotator as the correct annotation, the value of t can be decreased, possibly allowing for annotation instances with a smaller $M_{toSecond}$.

3.3 Machine learning system

As machine learning system, the conditional random fields system CRF++ (Kudo, 2013) will be

used. This system uses a combination of forward Viterbi and backward A* search for finding the best classification sequence for an input sentence, given the trained model. It can also produce the n-best classification sequences for each sentence, which is necessary for the proposed pre-tagger that presents the two best pre-taggings to the human annotator.

CRF++ can also give the conditional probably for the output, that is for the entire classification sequence of a sentence, which is needed in the proposed active learning algorithm.

3.4 Materials

There is a corpus of Swedish clinical text, i.e. the text in the narrative part of the health record, that contains clinical text from the Stockholm area, from the years 2006-2008 (Dalianis et al., 2009). A subset of this corpus, containing texts from an emergency unit of internal medicine, has been annotated for four types of named entities: *disorder*, *finding*, *pharmaceutical drug* and *body structure* (Skeppstedt et al., 2012). For approximately one third of this annotated corpus, double annotation has been performed, and the instances, for which there were a disagreement, have been resolved by one of the annotators.

The annotated corpus will form the main source of materials for the study proposed here, and additional data to annotate will be selected from a pool of unlabelled data from internal medicine emergency notes.

The larger subset of the annotated data, only annotated by one annotator, will be referred to as *Single* (containing 45 482 tokens), and the smaller subset, annotated by two annotators, will be referred to as *Double* (containing 25 370 tokens). The *Single* subset will be the main source for developing the pre-annotation/active learning method, whereas the *Double* subset will be used for a final evaluation.

3.5 Step-by-step explanation

The proposed method can be divided into 8 steps:

1. Train a CRF model with a randomly selected subset of the *Single* part of the annotated corpus, the *seed set*. The size of this *seed set*, as well as suitable features for the CRF model will be evaluated using cross validation on the *seed set*. The size should be as small as possible, limiting the amount of initial annotation needed, but large enough to have results in line with a baseline system using terminology matching for named entity recognition (Skeppstedt et al., 2012).

2. Apply the constructed CRF model on unlabelled data from the pool of data from internal medicine emergency notes. Let the model, which operates on a sentence level, provide the three most probable label sequences for each sentence, together with its level of certainty.

3. Calculate the difference in certainty between the most probable and the third most probable suggestion sequence for each sentence, that is $M_{toThird}$. Start with a low threshold t and place all sentences with $M_{toThird}$ above the threshold t in a list of candidates for presenting to the annotator (that is the sentences fulfilling the criterion $P(c_{p1}|\boldsymbol{x}) - P(c_{p3}|\boldsymbol{x}) > t$).

4. Order the sentences in the list of selected candidates in increasing order of $M_{toSecond}$. Present the sentence with the lowest $M_{toSecond}$ to the annotator. This is the sentence, for which the pre-tagger is most uncertain of which one of the two most probable pre-taggings is correct.

   Present the most probable pre-annotation as well as the second most probable pre-annotation, as shown in Figure 1.

5. If the annotator chooses that none of the presented pre-annotations is correct, discard the previous candidate selection and make a new one from the pool with a higher threshold value t. Again, order the sentences in increasing order of $M_{toSecond}$, and present the sentence with the lowest $M_{toSecond}$ to the annotator.

   Repeat step 3., 4. and 5., gradually increasing the threshold until the annotator accepts one of the presented pre-annotations.

6. Continue presenting the annotator with the two most probable pre-annotations for the sentences in the list of selected candidate sentences, and allow the human annotator to choose one of the pre-annotations.

The threshold t could be further adjusted according to how often the option 'None' is chosen.

7. Each selected annotation is added to a set of annotated data. When a sufficiently large amount of new sentences have been added to this set, the model needs to be retrained with the new data. The retraining of the model can be carried out as a background process while the human annotator is annotating. In order to use the annotator time efficiently, there should not be any waiting time while retraining.

8. When the model has been retrained, the process starts over from step 2.

3.6 Evaluation

The text passages chosen in the selection process will, as explained above, be used to re-train the machine learning model, and used when selecting new text passages for annotation. The effect of adding additional annotations will also be constantly measured, using cross validation on the *seed set*. The additional data added by the active learning experiments will, however, not be used in the validation part of the cross validation, but only be used as additional training data, in order to make sure that the results are not improved due to easily classified examples being added to the corpus.

When an actively selected corpus of the same size as the entire *Single* subset of the corpus has been created, this actively selected corpus will be used for training a machine learning model. The performance of this model will then be compared to a model trained on the single subset. Both models will be evaluated on the *Double* subset of the corpus. The hypothesis is that the machine learning model trained on the corpus partly created by pre-tagging and active learning will perform better than the model created on the original *Single* subset.

4 Conclusion

A method that combines pre-annotation and active learning, while reducing annotation bias, is proposed. A program for presenting pre-annotated data to the human annotator for selection has been constructed, and a corpus of annotated data suitable as a seed set and as evaluation data has been constructed. The active learning part of the proposed method remains, however, to be implemented.

Applying the proposed methods aims at creating a corpus suitable for training a machine learning system to recognise the four entities *Disorder*, *Finding*, *Pharmaceutical drug* and *Body structure*. Moreover, methods for facilitating annotated corpus construction will be explored, potentially adding new knowledge to the science of annotation.

Acknowledgements

I am very grateful to the reviewers and the pre-submission mentor for their many valuable comments. I would also like to thank Hercules Dalianis and Magnus Ahltorp as well as the participants of the 'Southern California Workshop on Medical Text Analysis and Visualization' for fruitful discussions on the proposed method.

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F 4th Styler, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, Wayne Ward, Martha Palmer, and Guergana K Savova. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc*, Jan.

Thorsten Brants and Oliver Plaehn. 2000. Interactive corpus annotation. In *LREC*. European Language Resources Association.

Wendy W Chapman, John N Dowling, and George Hripcsak. 2008. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. *Int J Med Inform, Epub 2007 Feb 20*, 77(2):107–113, February.

Wen-chi Chou, Richard Tzong-han Tsai, and Ying-shan Su. 2006. A semi-automatic method for annotating a biomedical proposition bank. In *FLAC'06. ACL*.

Hercules Dalianis, Martin Hassel, and Sumithra Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research, Kalmar, Sweden*, pages 243–249.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on pos-tagged corpus development.

In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 56–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rob Koeling, John Carroll, Rosemary Tate, and Amanda Nicholson. 2011. Annotating a corpus of clinical text records for learning to recognize symptoms automatically. In *Proceedings of the LOUHI 2011, Third International Workshop on Health Document Text Mining and Information Analysis*.

Taku Kudo. 2013. CRF++: Yet Another CRF toolkit. http://crfpp.sourceforge.net/. Accessed 2013-05-21.

Agnieszka Mykowiecka and Małgorzata Marciniak. 2011. Some remarks on automatic semantic annotation of a medical corpus. In *Proceedings of the LOUHI 2011, Third International Workshop on Health Document Text Mining and Information Analysis*.

Mariana Neves and Ulf Leser. 2012. A survey on annotation tools for the biomedical literature. *Briefings in Bioinformatics*.

Philip Ogren, Guergana Savova, and Christopher Chute. 2008. Constructing evaluation corpora for automated clinical named entity recognition. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 3143–3149, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Fredrik Olsson. 2008. *Bootstrapping Named Entity Annotation by Means of Active Machine Learning*. Ph.D. thesis, University of Gothenburg. Faculty of Arts.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *J. of Biomedical Informatics*, 42:950–966, October.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, IDA '01, pages 309–318, London, UK, UK. Springer-Verlag.

Andrew I. Schein and Lyle H. Ungar. 2007. Active learning for logistic regression: an evaluation. *Mach. Learn.*, 68(3):235–265, October.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1070–1079, Stroudsburg, PA, USA. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Maria Skeppstedt and Hercules Dalianis. 2012. Using active learning and pre-tagging for annotating clinical findings in health record text. In *Proceedings of SMBM 2012 - The 5th International Symposium on Semantic Mining in Biomedicine*, pages 98–99, Zurich, Switzerland, September 3-4.

Maria Skeppstedt, Maria Kvist, and Hercules Dalianis. 2012. Rule-based entity recognition and coverage of SNOMED CT in Swedish clinical text. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1250–1257, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Christopher T. Symons, Nagiza F. Samatova, Ramya Krishnamurthy, Byung H. Park, Tarik Umar, David Buttler, Terence Critchlow, and David Hysom. 2006. Multi-criterion active learning in conditional random fields. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '06, pages 323–331, Washington, DC, USA. IEEE Computer Society.

Katrin Tomanek, Philipp Daumke, Frank Enders, Jens Huber, Katharina Theres, and Marcel Müller. 2012. An interactive de-identification-system. In *Proceedings of SMBM 2012 - The 5th International Symposium on Semantic Mining in Biomedicine*, pages 82–86, Zurich, Switzerland, September 3-4.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, March.

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *J Am Med Inform Assoc*, 17(5):519–523.

Yefeng Wang. 2009. Annotating and recognising named entities in clinical notes. In *Proceedings of the ACL-IJCNLP Student Research Workshop*, pages 18–26, Singapore.

Fei Xia and Meliha Yetisgen-Yildiz. 2012. Clinical corpus annotation: Challenges and strategies. In *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop. Turkey.*