Using Soft Constraints in Joint Inference for Clinical Concept Recognition

Prateek Jindal and **Dan Roth**
Department of Computer Science, UIUC
201 N. Goodwin Ave, Urbana, IL 61801, USA
{jindal2, danr}@illinois.edu

Abstract

This paper introduces IQPs (Integer Quadratic Programs) as a way to model joint inference for the task of concept recognition in clinical domain. IQPs make it possible to easily incorporate soft constraints in the optimization framework and still support exact global inference. We show that soft constraints give statistically significant performance improvements when compared to hard constraints.

1 Introduction

In this paper, we study the problem of concept recognition in the clinical domain. State-of-the-art approaches (de Bruijn et al., 2011; Patrick et al., 2011; Torii et al., 2011; Minard et al., 2011; Jiang et al., 2011; Xu et al., 2012; Roberts and Harabagiu, 2011; Jindal and Roth, 2013) for concept recognition in clinical domain (Uzuner et al., 2011) use sequence-prediction models like CRF (Lafferty et al., 2001), MEMM (McCallum et al., 2000) etc. These approaches are limited by the fact that they can model only local dependencies (most often, first-order models like linear chain CRFs are used to allow tractable inference).

Clinical narratives, unlike newswire data, provide a domain with significant knowledge that can be exploited systematically to improve the accuracy of the prediction task. Knowledge in this domain can be thought of as belonging to two categories: (1) *Background Knowledge* captured in medical ontologies like UMLS (Url1, 2013), MeSH and SNOMED CT and (2) *Discourse Knowledge* driven by the fact that the narratives adhere to a specific writing style. While the former can be used by generating more expressive knowledge-rich features, the latter is more interesting from our current perspective, since it provides global constraints on what *output* structures are likely and what are not. We exploit this structural knowledge in our global inference formulation.

Integer Linear Programming (ILP) based approaches have been used for global inference in many works (Roth and Yih, 2004; Punyakanok et al., 2004; Punyakanok et al., 2008; Marciniak and Strube, 2005; Bramsen et al., 2006; Barzilay and Lapata, 2006; Riedel and Clarke, 2006; Clarke and Lapata, 2007; Clarke and Lapata, 2008; Denis et al., 2007; Chang et al., 2011). However, in most of these works, researchers have focussed only on hard constraints while formulating the inference problem.

Formulating all the constraints as hard constraints is not always desirable because the constraints are not perfect in many cases. In this paper, we propose Integer Quadratic Programs (IQPs) as a way of formulating the inference problem. IQPs is a richer family of models than ILPs and it enables us to easily incorporate soft constraints into the inference procedure. Our experimental results show that soft constraints indeed give much better performance than hard constraints.

2 Identifying Medical Concepts

Task Description Our input consists of clinical reports in free-text (unstructured) format. The task is: (1) to identify the boundaries of medical concepts and (2) to assign types to such concepts. Each concept can have 3 possible types: (1) Test, (2) Treatment, and (3) Problem. We would refer to these three types by `TEST`, `TRE` and `PROB` in the following discussion.

Our Approach In the first step, we identify the concept boundaries using a CRF (with BIO encod-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1808–1814,
Seattle, Washington, USA, 18-21 October 2013.

Test Problem Problem
[Chest x-ray] gave positive evidence for [atelectasis] and [sarcoidosis].

(a) Example 1

Problem Problem Problem Problem Problem Problem
No [hemoptysis], [hematemesis], [urgency], [abdominal pain], [black or tarry stools], [dysuria].

(b) Example 2

Figure 1: This figure motivates the global inference procedure we used. For discussion, please refer to §2.

ing). Features used by the CRF include the constituents given by MetaMap (Aronson and Lang, 2010; Url2, 2013), shallow parse constituents, surface form and part-of-speech (Url3, 2013) of words in a window of size 3. We also use conjunctions of the features.

After finding concept boundaries, we determine the probability distribution for each concept over 4 possible types (TEST, TRE, PROB or NULL). These probability distributions are found using a multi-class SVM classifier (Chang and Lin, 2011). Features used for training this classifier include concept tokens, full text of concept, bi-grams, headword, suffixes of headword, capitalization pattern, shallow parse constituent, Metamap type of concept, MetaMap type of headword, occurrence of concept in MeSH (Url4, 2013) and SNOMED CT (Url5, 2013), MeSH and SNOMED CT descriptors.

Inference Procedure: The final assignment of types to concepts is determined by an inference procedure. The basic principle behind our inference procedure is: "*Types of concepts which appear close to one another are often closely related. For some concepts, type can be determined with more confidence. And relations between concepts' types guide the inference procedure to determine the types of other concepts.*" We will now explain it in more detail with the help of examples. Figure 1 shows two sentences in which the concepts are shown in brackets and correct (gold) types of concepts are shown above them.

First, consider first and second concepts in Figure 1a. These concepts follow the pattern: *[Concept1] gave positive evidence for [Concept2]*. In clinical narratives, such a pattern strongly suggests that *Concept1* is of type TEST and *Concept2* is of type PROB. Table 1 shows additional such patterns. Next, consider different concepts in Figure 1b. All

	Pattern
1	using [TRE] for [PROB]
2	[TEST] showed [PROB]
3	Patient presents with [PROB] status post [TRE]
4	use [TRE] to correct [PROB]
5	[TEST] to rule out [PROB]
6	Unfortunately, [TRE] has caused [PROB]

Table 1: Some patterns that were used in constraints.

these concepts are separated by commas and hence, form a list. It is highly likely that such concepts should have the same type.

3 Modeling Global Inference

Inference is done at the level of sentences. Suppose there are m concepts in a sentence. Each of the m concepts has to be assigned one of the following types: TEST, TRE, PROB or NULL. To represent this as an inference problem, we define the indicator variables $x_{i,j}$ where i takes values from 1 to m (corresponding to concepts) and j takes values from 1 to 4 (corresponding to 4 possible types). $p_{i,j}$ refers to the probability that the i^{th} concept has type j.

We can now write the following optimization problem to find the optimal concept types:

$$\max_{x} \sum_{i=1}^{m} \sum_{j=1}^{4} x_{i,j} \cdot p_{i,j} \tag{1}$$

$$\text{subject to} \sum_{j=1}^{4} x_{i,j} = 1 \qquad \forall i \tag{2}$$

$$x_{i,j} \in \{0, 1\} \qquad \forall i, j \tag{3}$$

The objective function in Equation (1) expresses the fact that we want to maximize the expected number of correct predictions in each sentence. Equation (2) enforces the constraint that each concept has a unique type. We would refer to these as **Type-1** constraints.

3.1 Constraints Used

In this subsection, we will describe two additional types of constraints (**Type-2** and **Type-3**) that were added to the optimization procedure described above. Whereas **Type-1** constraints described above were formulated as *hard constraints*, **Type-2** and **Type-3** constraints are formulated as *soft constraints*.

3.1.1 Type-2 Constraints

Certain constructs like comma, conjunction, etc. suggest that the 2 concepts appearing in them should have the same type. Figure 1b shows an example of such a constraint. Suppose that there are n_2 such constraints. Also, assume that the l^{th} constraint says that the concepts $\mathcal{R}_l$ and $\mathcal{S}_l$ should have the same type. To model this, we define a variable w_l as follows:

$$w_l = \sum_{m=1}^{4} (x_{\mathcal{R}_l,m} - x_{\mathcal{S}_l,m})^2 \tag{4}$$

Now, if the concepts $\mathcal{R}_l$ and $\mathcal{S}_l$ have the same type, then w_l would be equal to 0; otherwise, w_l would be equal to 2. So, the l^{th} constraint can be enforced by subtracting $(\rho_2 \cdot \frac{w_l}{2})$ from the objective function given by Equation (1). Thus, a penalty of ρ_2 would be enforced iff this constraint is violated.

3.1.2 Type-3 Constraints

Some short patterns suggest possible types for the concepts which appear in them. Each such pattern, thus, enforces a constraint on the types of corresponding concepts. Figure 1a shows an example of such a constraint. Suppose that there are n_3 such constraints. Also, assume that the k^{th} constraint says that the concept $\mathcal{A}_{1,k}$ should have the type $\mathcal{B}_{1,k}$ and that the concept $\mathcal{A}_{2,k}$ should have the type $\mathcal{B}_{2,k}$. Equivalently, the k^{th} constraint can be written as follows in boolean algebra notation: $(x_{\mathcal{A}_{1,k},\mathcal{B}_{1,k}} = 1) \wedge (x_{\mathcal{A}_{2,k},\mathcal{B}_{2,k}} = 1)$. For the k^{th} constraint, we introduce one more variable $z_k \in \{0, 1\}$ which satisfies the following condition:

$$z_k = 1 \Leftrightarrow x_{\mathcal{A}_{1,k},\mathcal{B}_{1,k}} \wedge x_{\mathcal{A}_{2,k},\mathcal{B}_{2,k}} \tag{5}$$

Using boolean algebra, it is easy to show that Equation (5) can be reduced to a set of linear inequalities. Thus, we can incorporate the k^{th} constraint in the optimization problem by adding to it the constraint given by Equation (5) and by subtracting $(\rho_3(1 - z_k))$ from the objective function given by Equation (1). Thus, a penalty of ρ_3 is imposed iff k^{th} constraint is not satisfied ($z_k = 0$).

$$\max_x \sum_{i=1}^{m}\sum_{j=1}^{4} x_{i,j} \cdot p_{i,j} - \sum_{k=1}^{n_3} \rho_3(1 - z_k) - \sum_{l=1}^{n_2}\left(\rho_2 \cdot \frac{\sum_{m=1}^{4}(x_{\mathcal{R}_l,m} - x_{\mathcal{S}_l,m})^2}{2}\right) \tag{6}$$

$$\text{subject to} \sum_{j=1}^{4} x_{i,j} = 1 \qquad \forall i \tag{7}$$

$$x_{i,j} \in \{0, 1\} \qquad \forall i, j \tag{8}$$

$$z_k = 1 \Leftrightarrow x_{\mathcal{A}_{1,k},\mathcal{B}_{1,k}} \wedge x_{\mathcal{A}_{2,k},\mathcal{B}_{2,k}} \forall k \in \{1...n_3\} \tag{9}$$

Figure 2: Final Optimization Problem (an IQP)

3.2 Final Optimization Problem - An IQP

After incorporating all the constraints mentioned above, the final optimization problem (an IQP) is shown in Figure 2. We used Gurobi toolkit (Url6, 2013) to solve such IQPs. In our case, it solves 76 IQPs per second on a quad-core server with Intel Xeon X5650 @ 2.67 GHz processors and 50 GB RAM.

4 Experiments and Results

4.1 Datasets and Evaluation Metrics

For our experiments, we used the datasets provided by i2b2/VA team as part of 2010 i2b2/VA shared task (Uzuner et al., 2011). The datasets used for this shared task contained de-identied clinical reports from three medical institutions: Partners Healthcare (`PH`), Beth-Israel Deaconess Medical Center (`BIDMC`) and the University of Pittsburgh Medical Center (UPMC). UPMC data was divided into 2 sections, namely discharge (`UPMCD`) and progress notes (`UPMCP`). A total of 349 training reports and 477 test reports were made available to the participants. However, data which came from UPMC (more than 50% data) was not made available for public use. As a result, we had only 170 clinical reports for training and 256 clinical reports for testing. Table 3 shows the number of clinical reports made available by different institutions. The

	B			BK			BC			BKC		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
TEST	92.4	79.4	85.4	91.9	80.2	85.7	92.7	79.6	85.7	92.1	80.4	**85.8**
TRE	92.1	73.6	81.8	92.0	79.5	85.3	92.3	76.8	83.8	92.0	80.2	**85.7**
PROB	83.6	83.6	83.6	88.9	83.7	86.3	85.9	83.8	84.8	89.6	83.9	**86.7**
OVERALL	88.4	79.4	83.6	90.7	81.4	85.8	89.6	80.5	84.8	91.0	81.7	**86.1**

Table 2: Our final system, **BKC**, consistently performed the best among all 4 systems (**B**, **BK**, **BC** and **BKC**).

	PH	BIDMC	~~UPMCD~~	~~UPMCP~~
Train	97	73	~~98~~	~~81~~
Test	133	123	~~102~~	~~119~~

Table 3: Dataset Characteristics

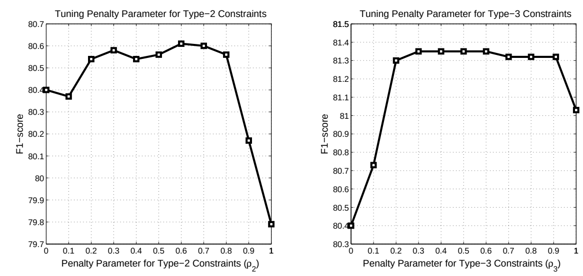
strikethrough text in this table indicates that the data was not made available for public use and hence, we couldn't use it. We used about 20% of the training data as a development set. For evaluation, we report precision, recall and F1 scores.

4.2 Results

In this section, we would refer to following 4 systems: (1) *Baseline* **(B)**, (2) *Baseline + Knowledge* **(BK)**, (3) *Baseline + Constraints* **(BC)** and (4) *Baseline + Knowledge + Constraints* **(BKC)**. Please note that the difference between **B** and **BK** is that **B** does not use the features derived from domain-specific knowledge sources (namely MetaMap, UMLS, MeSH and SNOMED CT) for training the classifiers. Both **B** and **BK** do not use the inference procedure. **BKC** uses all the features and also the inference procedure. In addition to these 4 systems, we would refer to another system, namely, **BKC-HARD**. This is similar to **BKC** system. However, it sets $\rho_2 = \rho_3 = 1$ which effectively turns **Type-2** and **Type-3** constraints into hard constraints by imposing very high penalty.

4.2.1 Importance of Soft Constraints

Figures 3a and 3b show the effect of varying the penalties (ρ_2 and ρ_3) for **Type-2** and **Type-3** constraints respectively. These figures show the F1-score of **BKC** on the development set. Penalty of 0 means that the constraint is not active. As we increase the penalty, the constraint becomes stronger. As the penalty becomes 1, the constraint becomes hard in the sense that final assignments must respect the constraint. We observe from Figures 3a and 3b that for **Type-2** and **Type-3** constraints, global maxima is attained at $\rho_2 = 0.6$ and $\rho_3 = 0.3$ respectively.



(a) Type-2 Constraints **(b)** Type-3 Constraints

Figure 3: These figures show the result of tuning the penalty parameters (ρ_2 and ρ_3) for soft constraints.

	BKC-HARD	**BKC**
TEST	84.7	**85.8**
TRE	84.7	**85.7**
PROB	85.6	**86.7**
OVERALL	85.1	**86.1**

Table 4: Soft constraints **(BKC)** consistently perform much better than hard constraints **(BKC-HARD)**.

Hard vs Soft Constraints Table 4 compares the performance of **BKC-HARD** with that of **BKC**. First 3 rows in this table show the performance of both systems for the individual categories (TEST, TRE and PROB). The fourth row shows the overall score of both systems. **BKC** outperformed **BKC-HARD** on all the categories by statistically significant differences at $p = 0.05$ according to Bootstrap Resampling Test (Koehn, 2004). For the OVERALL category, **BKC** improved over **BKC-HARD** by 1.0 F1 points.

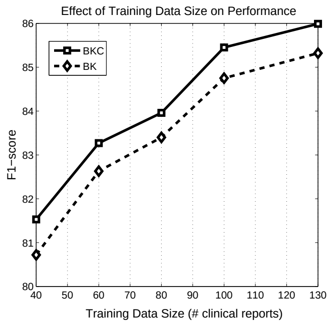


Figure 4: This figure shows the effect of training data size on performance of concept recognition.

4.2.2 Comparing with state-of-the-art baseline

In the 2010 i2b2/VA shared task, majority of top systems were CRF-based models, motivating the use of CRF as our baseline. Table 2 compares the performance of 4 systems: **B**, **BK**, **BC** and **BKC**. As pointed out before, our **BK** system uses CRF for boundary detection, employs all the knowledge-based features and is very similar to the top-performing systems in i2b2 challenge. We see from Table 2 that **BKC** consistently performed the best for individual as well as overall categories[1]. This result is statistically significant at $p = 0.05$ according to Bootstrap Resampling Test (Koehn, 2004). It should also be noted that **BC** performed significantly better than **B** for all the categories. Thus, the constraints are helpful even in the absence of knowledge-based features. Since we report results on publicly available datasets, future works would be able to compare their results with ours.

4.2.3 Effect of training data size

In Figure 4, we report the overall F1-score on a part of the development set as we vary the size of the training data from 40 documents to 130 documents. We notice that the performance increases steadily as more and more training data is provided. This suggests that if we could train on full training data as was made available in the challenge, the final scores could be much higher. We also notice from the figure that **BKC** consistently outperforms the state-of-the-art **BK** system as we vary the size of the training data, indicating the robustness of the joint inference procedure.

[1]Please note that the results reported in Table 2 can not be directly compared with those reported in the challenge because we only had a fraction of the original training and testing data.

5 Discussion and Related Work

In this paper, we chose to train a rather simple sequential model (using CRF), and focused on incorporating global constraints only at inference time[2]. While it is possible to jointly train the model with the global constraints (as illustrated by Chang et al. (2007), Mann and McCallum (2007), Mann and McCallum (2008), Ganchev et al. (2010) etc.), this process will be a lot less efficient, and prior work (Roth and Yih, 2005) has shown that it may not be beneficial.

Roth and Yih (2004, 2007) suggested the use of integer programs to model joint inference in a fully supervised setting. Our paper follows their conceptual approach. However, they used only hard constraints in their inference formulation. Chang et al. (2012) extended the ILP formulation and used soft constraints within the Constrained Conditional Model formulation (Chang, 2011). However, their implementation performed only approximate inference. In this paper, we extended the integer linear programming to a quadratic formulation, arguing that it simplifies the modeling step[3], and showed that it is possible to do exact inference efficiently.

Conclusion

This paper presented a global inference strategy (using IQP) for concept recognition which allows us to model structural knowledge of the clinical domain as soft constraints in the optimization framework. Our results showed that soft constraints are more effective than hard constraints.

Acknowledgments

This research was supported by Grant HHS 90TR0003/01 and by IARPA FUSE program via DoI/NBC contract #D11PC2015. Its contents are solely the responsibility of the authors and do not necessarily represent the official views, either expressed or implied, of the HHS, IARPA, DoI/NBC or the US government. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

[2]In another experiment, we replaced the CRF with an MEMM. Surprisingly, MEMM performed as well as CRF.

[3]It should be noted that it is possible to reduce IQPs to ILPs using variable substitution. However, the resulting ILPs can be exponentially larger than original IQPs.

References

A.R. Aronson and F.M. Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229.

R. Barzilay and M. Lapata. 2006. Aggregation via set partitioning for natural language generation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 359–366. Association for Computational Linguistics.

P. Bramsen, P. Deshpande, Y.K. Lee, and R. Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 189–198. Association for Computational Linguistics.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

M. Chang, L. Ratinov, and D. Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *Association for Computational Linguistics*, pages 280–287, Prague, Czech Republic, 6. Association for Computational Linguistics.

K.-W. Chang, R. Samdani, A. Rozovskaya, N. Rizzolo, M. Sammons, and D. Roth. 2011. Inference protocols for coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 40–44, Portland, Oregon, USA. Association for Computational Linguistics.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2012. Structured learning with constrained conditional models. *Machine learning*, pages 1–33.

M. Chang. 2011. *Structured Prediction with Indirect Supervision*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

James Clarke and Mirella Lapata. 2007. Modelling compression with discourse constraints. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1–11.

J. Clarke and M. Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31(1):399–429.

B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562.

P. Denis, J. Baldridge, et al. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–243.

Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049.

M. Jiang, Y. Chen, M. Liu, S.T. Rosenbloom, S. Mani, J.C. Denny, and H. Xu. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Info Assoc*, 18(5):601–606.

P. Jindal and D. Roth. 2013. End-to-end coreference resolution for clinical narratives. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2106–2112, 8.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of Empirical Methods in Natural Language Processing*, volume 4, pages 388–395.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gideon S Mann and Andrew McCallum. 2007. Simple, robust, scalable semi-supervised learning via expectation regularization. In *Proceedings of the 24th international conference on Machine learning*, pages 593–600. ACM.

Gideon Mann and Andrew McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Proceedings of Association for Computational Linguistics*, pages 870–878.

T. Marciniak and M. Strube. 2005. Beyond the pipeline: Discrete optimization in nlp. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 136–143. Association for Computational Linguistics.

A. McCallum, D. Freitag, and F. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 591–598.

A.L. Minard, A.L. Ligozat, A.B. Abacha, D. Bernhard, B. Cartoni, L. Deléger, B. Grau, S. Rosset, P. Zweigenbaum, and C. Grouin. 2011. Hybrid methods for improving information access in clinical documents: Concept, assertion, and relation identification. *J Am Med Info Assoc*, 18(5):588–593.

J.D. Patrick, D.H.M. Nguyen, Y. Wang, and M. Li. 2011. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association*, 18(5):574–579.

V. Punyakanok, D. Roth, W. Yih, and D. Zimak. 2004. Semantic role labeling via integer linear programming inference. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1346. Association for Computational Linguistics.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287.

S. Riedel and J. Clarke. 2006. Incremental integer linear programming for non-projective dependency parsing. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 129–137. Association for Computational Linguistics.

K. Roberts and S.M. Harabagiu. 2011. A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, 18(5):568–573.

D. Roth and W. Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of conference on Computational Natural Language Learning (CoNLL)*, pages 1–8. Association for Computational Linguistics.

D. Roth and W. Yih. 2005. Integer linear programming inference for conditional random fields. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 737–744.

D. Roth and W. Yih. 2007. Global inference for entity and relation identification via a linear programming formulation. *Introduction to Statistical Relational Learning*, pages 553–580.

M. Torii, K. Wagholikar, and H. Liu. 2011. Using machine learning for concept extraction on clinical documents from multiple data sources. *Journal of the American Medical Informatics Association*, 18(5):580–587.

Url1. 2013. Umls: Unified medical language system (http://www.nlm.nih.gov/research/umls/) (accessed july 1, 2013).

Url2. 2013. Metamap (http://metamap.nlm.nih.gov/) (accessed july 1, 2013).

Url3. 2013. Illinois part-of-speech tagger (http://cogcomp.cs.illinois.edu/page/software_view/pos) (accessed july 1, 2013).

Url4. 2013. Mesh: Medical subject headings (http://www.nlm.nih.gov/mesh/meshhome.html) (accessed july 1, 2013).

Url5. 2013. Snomed ct: Snomed clinical terms (http://www.ihtsdo.org/snomed-ct/) (accessed july 1, 2013).

Url6. 2013. Gurobi optimization toolkit (http://www.gurobi.com/) (accessed july 1, 2013).

O. Uzuner, B.R. South, S. Shen, and S.L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of American Medical Informatics Association*.

Y. Xu, K. Hong, J. Tsujii, I. Eric, and C. Chang. 2012. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5):824–832.