

Recognizing Disjoint Clinical Concepts in Clinical Text Using Machine

Learning-based Methods

Buzhou Tang, PhD^{1,2}, Qingcai Chen, PhD¹, Xiaolong Wang, PhD¹, Yonghui Wu, PhD²,
Yaoyun Zhang PhD², Min Jiang, MS², Jingqi Wang, MS², Hua Xu, PhD²

¹Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of
Technology Shenzhen Graduate School, Shenzhen, Guangdong, China

²School of Biomedical Informatics, The University of Texas Health Science Center at
Houston, Houston, Texas, USA

ABSTRACT

Clinical concept recognition (CCR) is a fundamental task in clinical natural language processing (NLP) field. Almost all current machine learning-based CCR systems can only recognize clinical concepts of consecutive words (called consecutive clinical concepts, CCCs), but can do nothing about clinical concepts of disjoint words (called disjoint clinical concepts, DCCs), which widely exist in clinical text. In this paper, we proposed two novel types of representations for disjoint clinical concepts, and applied two state-of-the-art machine learning methods to recognizing consecutive and disjoint concepts. Experiments conducted on the 2013 ShARe/CLEF challenge corpus showed that our best system achieved a “strict” F-measure of 0.803 for CCCs, a “strict” F-measure of 0.477 for DCCs, and a “strict” F-measure of 0.783 for all clinical concepts, significantly higher than the baseline systems by 4.2% and 4.1% respectively.

INTRODUCTION

With rapid growth of medical informatics technology, a large number of electronic health records (EHRs) have been available in recent years, including a huge mass of data, such as clinical narratives. They have been being used not only to support computerized clinical systems (e.g., computerized clinical decision support systems [1][2]), but also to help the development of clinical and translational research [3]. One of the challenges to use them is that much information is embedded in clinical notes, but cannot be directly accessible for computerized clinical systems which rely on structured information. Therefore, natural language processing (NLP) technologies, which can extract structured information from narrative text, have received great attention in medical domain [4], and many clinical NLP systems have been developed for different applications [5].

Clinical concept recognition (CCR) as a fundamental task of clinical NLP has also attracted great attention, and a large number of systems have been developed to recognize clinical concepts from various types of clinical notes in last two decades. Earlier systems were based on rules or dictionaries manually built by medical experts. The representative systems included MedLEE [4], SymText/MPlus [6][7], MetaMap [8], KnowledgeMap [9], cTAKES [10], and HiTEX [11]. In the past few years, with the increasingly available annotated clinical corpora, researchers started to apply machine learning algorithms to CCR, and several clinical NLP challenges were organized to promote the research on this task. For example, the Center for Informatics for Integrating Biology & the Beside (i2b2) organized NLP

challenges on CCR in 2009 and 2010 respectively [12][13]. The ShARe/CLEF eHealth Evaluation Lab (SHEL) organized a NLP challenge in 2013 [14], which includes a subtask of disorder recognition [15], and launched a similar NLP challenge as a part of the SemEval (Semantic Evaluation) in 2014 (i.e., SemEval-2014 Task 7) using the same training set, but different test set. Rule-based, machine learning-based and hybrid methods were developed to recognize medications by over twenty participating teams in the 2009 i2b2 NLP challenge [12]. In the 2010 i2b2 NLP challenge, clinical concepts, including problems, tests and treatments, not limited to medication, were required to recognize [13]. Most systems were primarily based on machine learning algorithms in this challenge, likely due to a large available annotated corpus [14]. In both the 2013 ShARe/CLEF and SemEval-2014 challenges, machine learning-based systems achieved state-of-the-art performance on disorder concept recognition [15]. Among the four NLP challenges, the ShARe/CLEF and SemEval-2014 challenges first considered disjoint clinical concepts (DCCs), which consist of multiple non-consecutive sequences of tokens. Actually, DCCs always drew much attention because they widely exists in clinical text and are important for subsequent applications such as clinical reasoning systems. In the 2013 ShARe/CLEF challenge, CCR is a preliminary step for concept mapping, mapping concepts in clinical text to UMLS concepts. Almost no machine learning-based method was proposed for disjoint clinical concept recognition except some simple rule-based systems such as MetaMap [8] and cTAKES [10], as well as disjoint named entity recognition in other domains. The main reason may lie in that it is more difficult to annotate DCCs than CCCs. In the 2013 ShARe/CLEF challenge, the organizers annotated a corpus with both CCCs and DCCs where DCCs account for about 10%. This corpus gave us a good chance to investigate how to recognize DCCs using machine learning-based methods. Compared with CCC recognition, the main challenge of DCC recognition is how to represent them. In previous studies, clinical concepts were typically represented by “BIO” tags, where B, I and O denotes that a token is at the beginning, inside and outside of a concept respectively [16]. This representation worked very well for CCCs, but not suitable for DCCs. For DCC recognition, a few systems tried some methods in the 2013 ShARe/CLEF challenge, including rule-based [17] and machine learning-based [18][19] [20] methods. The machine learning-based methods showed much better performance than the rule-based methods. Among all machine learning-based methods, our method was the best [14], which proposed a novel representation for DCCs.

In this paper, we proposed another type of representation for DCCs based on our previous work for the 2013 ShARe/CLEF challenge, which ranked first among 20 participating teams [14][15] on the disorder recognition task. In that study, we proposed a novel type of representation for DCCs by using two additional tags (i.e., H-head entity and D-non-head entity) based on the representation for CCCs, and a two-step method to recognize them. Although this type of representation can completely separate DCCs from CCCs, it does not provide enough information about how to combine head entities and non-head entities except that head entities should be shared with more than one DCC, and non-head entities should be combined with other head/non-head entities. Therefore, Some extra rules is needed for head/non-head entity combination. In this study, we proposed another type of representation for DCCs that integrates combination information into the representation strategy. Using this representation, a separate step for combination is not required any more. Similar to our

previous work in [20], we compared Conditional Random Fields (CRF) [21] and Structured Support Vector Machines (SSVM) [22] when using the two types of representations. To prove the effectiveness of our methods, we also compared them with the CRF/SSVM-based systems ignoring DCCs. To the best of our knowledge, it is the first time to comprehensively investigate DCC recognition using machine learning-based methods, which can be used as a benchmark for further studies. Moreover, the methods proposed in this paper can also be easily applicable to recognize disjoint named entities in other domains.

METHODS

Representations for CCCs and DCCs

We adopted “BIO” tags to represent CCCs in the whole paper. For convenience, we did not repeat any more. For DCCs, two different types of representations were proposed. The first representation used “BIOHD”, where ‘H’ denotes head entities which are consecutive sequences of tokens shared by multiple disjoint concepts in a sentence, and ‘D’ denotes non-head entities which are consecutive sequences of tokens in a disjoint concept not shared by other disjoint concepts in a sentence. The second representation used “BIOHD1234”, where ‘1’, ‘2’, ‘3’ and ‘4’ denote that a non-head entity is combined with the nearest head entity at left, the nearest non-head entity at left, the nearest head entity at right and the nearest non-head entity at right respectively. Figure 1 shows some examples of DCCs represented by “BIOHD” and “BIOHD1234”, where three sentences were used to illustrate different cases of DCCs. In sentence 1, two disjoint disorders (i.e., “extremities ... turned in” and “extremities ... clinched together”) share a head entity (i.e., “extremities”). The head entity is represented by “extremities/HB-disorder” no matter using “BIOHD” or “BIOHD1234”. The non-head entities (i.e., “turned in” and “clinched together”) are represented by “turned/DB-disorder in/DI-disorder” and “clinched/DB-disorder together/DI-disorder” when using “BIOHD”, while are represented by “turned/D1B-disorder in/D1I-disorder” and “clinched/D1B-disorder together/D1I-disorder” when using “BIOHD1234” as both of them should be combined with the nearest head entity at left. Considering that head entities are represented by “H” in both “BIOHD” and “BIOHD1234”, and non-head entities are always represented by “D” when using “BIOHD”, we did not explain them any more in the following examples. All non-head entities mentioned in the following examples are the cases of representing DCCs using “BIOHD1234”. In sentence 2, two disjoint disorders (i.e., “ABD ... distend” and “ABD ... tenderness ... RUQ”) share a head entity (i.e., “ABD”). The non-head entity in the first disjoint disorder (i.e., “distend”) is represented by “D1B-disorder” as it should be combined with the nearest head entity at left. The non-head entities (i.e., “tenderness” and “RUQ”) in the second disjoint disorder are represented by “D1B-disorder” and “D2B-disorder” respectively because “tenderness” should be combined with the nearest head entity at left, while “RUQ” should be combined with the nearest non-head entity at left (i.e., “tenderness”). In sentence 3, there is a disjoint disorder (i.e., “left atrium ... dilated”) composed of two non-head entities (i.e., “left atrium” and “dilated”), which is represented by “D4D2...D2” (i.e., “left/D4B-disorder atrium/D4I-disorder” and “dilated/D2B-disorder”) as the first non-head entity (at left) should be combined with the second non-head entity (at right), and vice versa.

Sentence 1: “Her **extremities** appeared **turned in** and **clinched together** except for her right arm .”
 Concepts: “extremities ... turned in” and “extremities ... clinched together”
 BIOHD: “Her/O extremities/HB-disorder appeared/O turned/DB-disorder in/DI-disorder and/O clinched/DB-disorder together/DI-disorder except/O for/O her/O right/O arm/O ./O”
 BIOHD1234: “Her/O extremities/HB-Disorder appeared/O turned/D1B-disorder in/D1I-disorder and/O clinched/D1B-disorder together/D1I-disorder except/O for/O her/O right/O arm/O ./O”

Sentence 2: “**ABD** : mild **distend** , mild **tenderness** in **RUQ** .”
 Concepts: “ABD ... distend” and “ABD ... tenderness ... RUQ”
 BIOHD: “ABD/HB-disorder ./O mild/O distend/DB-disorder ./O mild/O tenderness/DB-disorder in/O RUQ/DB-disorder ./O”
 BIOHD1234: “ABD/HB-disorder ./O mild/O distend/D1B-disorder ./O mild/O tenderness/D1B-disorder in/O RUQ/D2B-disorder ./O”

Sentence 3: “The **left atrium** is **dilated** .”
 Concepts: “left atrium ... dilated”
 BIOHD: “The/O left/DB-disorder atrium/DI-disorder is/O dilated/DB-disorder ./O”
 BIOHD1234: “The/O left/D4B-disorder atrium/D4I-disorder is/O dilated/D2B-disorder ./O”

Figure 1. Examples of two different representations for disjoint clinical concepts.

When using ”BIOHD” tags to represent disjoint clinical concepts, it is required to develop another system to determine how to combine head entities and non-head entities in a sentence as mentioned in our previous work [20], where two simple rules are proposed for combination as follows:

- 1) For each head entity, combine it with all other non-head entities to form disjoint concepts.
- 2) If no head entity, combine all non-head entities together to form a disjoint concept.

When using “BIOHD1234” tags to represent DCCs, it is straight forward to combine head entities and non-head entities since the combination information has been integrated into tags. For example, in the sentence 2 in Figure 1, there are two paths from non-head entities to the head entities: 1) distend→ABD; 2) RUQ→tenderness→ABD, forming two disjoint disorders.

Machine learning-based consecutive and disjoint CCR

When CCCs and DCCs are represented by “BIOHD” or “BIOHD1234”, CCR can be recognized as a classification problem, where each token of a sentence is labeled with a tag of “BIOHD” or “BIOHD1234” as shown in Figure 1. For “BIOHD”, the tag set is {B-disorder, I-disorder, O, HB-disorder, HI-disorder, DB-disorder, DI-disorder}. For “BIOHD1234”, the tag set is {B-disorder, I-disorder, O, HB-disorder, D1B-disorder, D1I-disorder, D2B-disorder, D2I-disorder, D3B-disorder, D3I-disorder, D4B-disorder, D4I-disorder}. Moreover, the machine learning methods proposed for consecutive CCR can also be applied to consecutive and disjoint CCR, such as Support Vector Machine (SVM) [13], CRF [23] , SSVM [24][25] and so on. In our study, we compared CRF and SSVM, two state-of-the-art machine learning methods for consecutive and disjoint CCR [25], and used CRFsuite (<http://www.chokkan.org/software/crfsuite/>) and SVM^{hmm} (http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html) as implements of CRF and SSVM respectively. For sentence boundary detection and tokenization, we adopted the corresponding modules of MedEx (<https://code.google.com/p/medex-uima/downloads/list>), a specific tool for medical information extraction.

Features

We used the same features as our previous work in [20], including bag-of-word, part-of-speech (POS), note type, section information, word representations, semantic categories of words. Most of them were also used in our previous systems for medical concept recognition [23][25][24][26]. For detailed information, please refer to the references.

Dataset

We used the dataset of the 2013 ShARe/CLEF challenge, which is composed of 298 notes from different clinical encounters including radiology reports, discharge summaries, ECG reports and ECHO reports. For each note, only disorders, including consecutive and disjoint disorders, were annotated according to a pre-defined guideline. The dataset was divided into two parts: a training set of 199 notes used for system development, and a test set of 99 notes used for system evaluation. In the training set, 651 out of 5811 disorders were disjoint, and 439 out of 5340 disorders were disjoint in the test set. Table 1 shows the counts of disorders in the training and test sets, where the number in parenthesis is the count of disjoint disorders.

Table 1. Counts of disorders in the training and test sets.

Dataset	Type	#Note	#Concept	#Consecutive	#Disjoint
Training	All	199	5816	5165	651
	ECHO	42	828	603	225
	RADIOLOGY	42	555	489	66
	DISCHARGE	61	3589	3232	357
	ECG	54	193	190	3
Test	All	99	5340	4901	439
	ECHO	12	338	280	58
	RADIOLOGY	12	162	158	4
	DISCHARGE	75	4840	4463	377
	ECG	0	0 (0)	0	0

Experiments and Evaluation

All models were trained on the training set and evaluated on the test set, and their parameters were optimized by 10-fold cross-validation on the training set. The performance of disorder concept recognition were evaluated by precision (P), recall (R) and F-measure (F) in both “strict” and “relaxed” modes [14], calculated by an evaluation tool provided by the organizers (<https://sites.google.com/site/shareclefehealth/evaluation>). To investigate the effect of the proposed methods, we started with two types of baseline systems: one treated every head and non-head entity of disjoint clinical concepts as an individual CCC (1st); and the other one removed all DCCs (2nd), and then compared them with our systems.

RESULTS

Table 2 shows the overall performance of the machine learning-based consecutive and disjoint CCR systems on the test set, when using different types of representations for CCCs and DCCs. The systems using “BIOHD” or “BIOHD1234” showed significantly better performance than the corresponding baseline systems, indicating that the proposed representations are suitable for CCCs and DCCs. For example, when using “BIOHD”, the CRF-based system outperformed the 1st baseline system by 3.9% (0.777 vs 0.738), and the 2nd baseline system by 5.9% (0.777 vs 0.718) in “strict” F-measure respectively; The SSVM-based system outperformed the 1st and 2nd baseline systems by 4.1% (0.782 vs 0.741) and 4.0% (0.782 vs 0.742) in “strict” F-measure respectively. When using the same representations for CCCs and DCCs, the SSVM-based systems showed better performance than the CRF-based systems. For example, the “strict” F-measure of the SSVM-based system was 0.782, while that of the CRF-based system was 0.777 when using “BIOHD”. For each machine learning method, the system using “BIOHD1234” slightly outperformed the system using “BIOHD”. For example, the SSVM-based system using “BIOHD1234” achieved a “strict” F-measure of 0.783, while the SSVM-based system using “BIOHD” achieved a “strict” F-measure of 0.782.

Here, we should note that the CRF-based and SSVM-based systems using “BIOHD” in this paper achieved better performance than our previous systems submitted to the 2013 ShARe/CLEF challenge as some bugs in previous systems have been fixed in current ones.

Table 2. Overall performance of the machine learning-based CCR systems.

System		Strict			Relaxed		
		P	R	F	P	R	F
1st Baseline	CRF	0.773	0.707	0.738	0.937	0.848	0.890
	SSVM	0.764	0.720	0.741	0.933	0.863	0.897
2nd Baseline	CRF	0.862	0.615	0.718	0.965	0.693	0.807
	SSVM	0.842	0.663	0.742	0.947	0.749	0.836
BIOHD	CRF	0.839	0.723	0.777	0.952	0.835	0.890
	SSVM	0.830	0.740	0.782	0.941	0.849	0.892
BIOHD1234	CRF	0.845	0.722	0.778	0.955	0.831	0.889
	SSVM	0.834	0.739	0.783	0.942	0.846	0.892

Furthermore, table 3 shows the performance of machine learning-based systems in the “strict” mode for CCCs and DCCs respectively. The machine learning-based system using “BIOHD” or “BIOHD1234” not only correctly recognized a number of DCCs, but also improved the performance for CCCs. For example, the “strict” F-measure of the CRF-based system using “BIOHD” for DCCs was 0.433, and that for CCCs was 0.799, higher than the 1st and 2nd baseline systems by 2.8% (0.799 vs 0.771) and 4.4% (0.799 vs 0.753). When using the same

representations for CCCs and DCCs, the SSVM-based systems showed better performance than CRF-based systems for both CCCs and DCCs. For example, when using “BIOHD”, the SSVM-based system achieved a “strict” F-measure of 0.802 for CCCs and a “strict” F-measure of 0.487 for DCCs; while the CRF-based system achieved a “strict” F-measures of 0.799 and 0.433 for CCCs and DCCs respectively. For each machine learning method, the systems using “BIOHD1234” showed slightly better performance than the systems using “BIOHD” for CCCs, but slightly worse performance for DCCs. For each machine learning-based system, the performance for CCCs is much higher than that for DCCs.

Table 3. Performance of the machine learning-based systems for CCCs and DCCs (“strict”).

System		CCCs			DCCs		
		P	R	F	P	R	F
1st	CRF	0.773	0.770	0.771	0.000	0.000	0.000
Baseline	SSVM	0.764	0.784	0.774	0.000	0.000	0.000
2nd	CRF	0.862	0.670	0.753	0.000	0.000	0.000
Baseline	SSVM	0.842	0.722	0.777	0.000	0.000	0.000
BIOHD	CRF	0.844	0.759	0.799	0.726	0.308	0.433
	SSVM	0.832	0.774	0.802	0.794	0.352	0.487
BIOHD	CRF	0.849	0.759	0.802	0.751	0.297	0.426
1234	SSVM	0.834	0.775	0.803	0.813	0.338	0.477

DISCUSSION

It is reasonable that the proposed two types of representations improved the performance of CCR systems as expected since they can distinguish DCCs from CCCs. To evaluate their representation abilities, we calculated upper boundaries of systems using “BIO”, “BIOHD” and “BIOHD1234” on the training set respectively. Firstly, we represented notes with gold clinical concepts using one of the three types of representations, and then converted the labeled notes back to clinical concepts. Finally, the upper boundary was calculated by comparing the converted clinical concepts with the gold ones using the evaluation tool. The upper boundaries of the systems using the three types of representations are shown in Table 3. The upper boundaries of systems using our representations were much higher than the systems using “BIO” (i.e., baseline systems shown in figure 1), and the differences ranged from 6.7% to 12.9% in “strict” F-measure. Among the proposed two types of representations, the upper boundary of the systems using “BIOHD1234” is slightly higher than the systems using “BIOHD” (0.969 vs 0.965 in “strict” F-measure). It means that the representation ability of “BIOHD1234” is slightly stronger than “BIOHD”, and is much stronger than “BIO”.

Although both “BIOHD” and “BIOHD1234” have good ability to represent both CCCs and DCCs, they are not complete. When clinical concepts are represented by “BIOHD”, clinical

Table 3. Upper boundaries of the systems using the three representations on the training set.

Representation	Strict			Relaxed		
	P	R	F	P	R	F
1st “BIO”	0.806	0.877	0.840	1.000	1.000	1.000
2nd “BIO”	0.999	0.816	0.898	1.000	0.817	0.899
“BIOHD”	0.967	0.962	0.965	1.000	0.999	0.999
“BIOHD1234”	0.976	0.963	0.969	1.000	0.997	0.999

concepts in the following three cases might be wrongly represented: 1) more than two head entities in a sentence; 2) more than one non-head entity in DCCs; 3) more than one DCC in a sentence. Some examples of clinical concepts wrongly represented by “BIOHD” are shown in figure 2. When clinical concepts are represented by “BIOHD1234”, the clinical concepts in sentence 2 and 3 are correctly represented, but the clinical concepts in sentence 1 are wrongly represented. Therefore, designing a complete representation for CCCs and DCCs is still a valuable topic, which is one case of our future work.

Sentence 1: “There is a small amount of **blood** seen within the **third** and **fourth ventricles** .”
 Gold: “blood ... third ventricles” and “blood ... four ... ventricles”
 BIOHD: “There/O is/O a/O small/O amount/O of/O blood/HB-disorder seen/O within/O the/O third/DB-disorder and/O fourth/DB-disorder ventricles/HB-disorder ./O”
 Converted: “blood ... third”, “blood ... fourth”, “third ... ventricles” and “fourth ... ventricles”

Sentence 2: “**ABD** : mild **distend** , mild **tenderness** in **RUQ** .”
 Gold: “ABD ... distend” and “ABD ... tenderness ... RUQ”
 BIOHD: “ABD/HB-disorder ./O mild/O distend/DB-disorder ./O mild/O tenderness/DB-disorder in/O RUQ/DB-disorder ./O”
 Converted: “ABD ... distend”, “ABD ... tenderness” and “ABD ... RUQ”

Sentence 3: “The patient had **blood** in his mouth and **on his tongue** , **pupils** were pinpoint and **reactive** .”
 Gold: “blood ... on his tongue” and “pupils ... reactive”
 BIOHD: “The/O patient/O had/O blood/DB-disorder in/O his/O mouth/O and/O on/DB-disorder his/DI-disorder tongue/DI-disorder ./O pupils/DB-disorder were/O pinpoint/O and/O reactive/DB-disorder ./O”
 Converted: “blood ... on his tongue ... pupils ... reactive”

Figure 2. Examples of clinical concepts wrongly represented by “BIOHD”.

Since the ratio of CCCs to DCCs in the 2103 ShARE/CLEF challenge corpus is about 9:1, which is imbalanced, it is easy to understand that the performance of machine learning-based systems for DCCs is not as good as the performance for CCCs. An interesting direction is to consider data imbalance for further improvement, which is another case of our future work.

Moreover, all methods proposed in this paper are not limited to clinical concepts, but also suitable for general named entities. Therefore, they are potentially useful in other domains.

CONCLUSIONS

In this study, we investigated the machine learning-based methods to recognizing CCCs and DCCs. Two novel types of representations were proposed for CCCs and DCCs, and their

effectiveness was proved on a benchmark dataset.

Acknowledgements

This study is supported in part by grants: National 863 Program of China (2015AA015405), NSFCs (National Natural Science Foundation of China) (61402128, 61473101, 61173075 and 61272383) and Strategic Emerging Industry Development Special Funds of Shenzhen (ZDSY20120613125401420, JCYJ20140508161040764, JCYJ20140417172417105 and JCYJ20140627163809422).

References

- [1] D. W. Bates, G. J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, C. Spurr, R. Khorasani, M. Tanasijevic, and B. Middleton, “Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality,” *J Am Med Inform Assoc*, vol. 10, no. 6, pp. 523–530, 2003.
- [2] K. B. Wagholikar, K. L. MacLaughlin, M. R. Henry, R. A. Greenes, R. A. Hankey, H. Liu, and R. Chaudhry, “Clinical decision support with automated text processing for cervical cancer screening,” *J Am Med Inform Assoc*, pp. amiajnl–2012–000820, Apr. 2012.
- [3] P. J. Embi and P. R. O. Payne, “Clinical Research Informatics: Challenges, Opportunities and Definition for an Emerging Domain,” *J Am Med Inform Assoc*, vol. 16, no. 3, pp. 316–327, May 2009.
- [4] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson, “A general natural-language text processor for clinical radiology,” *J Am Med Inform Assoc*, vol. 1, no. 2, pp. 161–174, Apr. 1994.
- [5] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, “Extracting information from textual documents in the electronic health record: a review of recent research,” *Yearb Med Inform*, pp. 128–144, 2008.
- [6] S. B. Koehler, “Symtext: a natural language understanding system for encoding free text medical data,” The University of Utah, 1998.
- [7] L. Christensen, P. Haug, and M. Fiszman, “Mplus: a probabilistic medical language understanding system,” 2002.
- [8] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: historical perspective and recent advances,” *J Am Med Inform Assoc*, vol. 17, no. 3, pp. 229–236, Jun. 2010.
- [9] J. C. Denny, P. R. Irani, F. H. Wehbe, J. D. Smithers, and A. Spickard, “The KnowledgeMap Project: Development of a Concept-Based Medical School Curriculum Database,” *AMIA Annu Symp Proc*, vol. 2003, pp. 195–199, 2003.
- [10] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,” *J Am Med Inform Assoc*, vol. 17, no. 5, pp. 507–513, 2010.
- [11] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, “Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system,” *BMC Medical Informatics and Decision Making*, vol. 6, no. 1, p. 30, Jul. 2006.
- [12] Ö. Uzuner, I. Solti, and E. Cadag, “Extracting medication information from clinical text,”

- J Am Med Inform Assoc*, vol. 17, no. 5, pp. 514–518, Sep. 2010.
- [13] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, “2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text,” *J Am Med Inform Assoc*, vol. 18, no. 5, pp. 552–556, Oct. 2011.
- [14] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, and N. Elhadad, “Overview of the ShARe/CLEF eHealth Evaluation Lab 2013,” presented at the CLEF 2013 Evaluation Labs and Workshop, 2013.
- [15] W. W. Chapman, G. Savova, and N. Elhadad, “ShARe/CLEF Shared Task 1 for boundary detection and normalization of SNOMED disorders,” presented at the CLEF 2013, 2013, p. to appear.
- [16] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, Jan. 2007.
- [17] Y. Xia, X. Zhong, P. Liu, C. Tan, S. Na, Q. Hu, and Y. Huang, “Combining MetaMap and cTAKES in Disorder Recognition: THCIB at CLEF eHealth Lab 2013 Task 1,” presented at the CLEF 2013 Evaluation Labs and Workshop, 2013.
- [18] J. Cogley, N. Stokes, and J. Carthy, “Medical Disorder Recognition with Structural Support Vector Machines,” presented at the CLEF 2013 Evaluation Labs and Workshop, 2013.
- [19] A. Bodnari, L. Deléger, T. Laverigne, A. Névéol, and P. Zweigenbaum, “A Supervised Named-Entity Extraction System for Medical Text,” presented at the CLEF 2013 Evaluation Labs and Workshop, 2013.
- [20] B. Tang, Y. Wu, M. Jiang, J. C. Denny, and H. Xu, “Recognizing and Encoding Disorder Concepts in Clinical Text using Machine Learning and Vector Space Model,” presented at the CLEF 2013, 2013, p. to appear.
- [21] J. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” *Departmental Papers (CIS)*, 2001.
- [22] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, “Large margin methods for structured and interdependent output variables,” *JOURNAL OF MACHINE LEARNING RESEARCH*, vol. 6, pp. 1453–1484, 2005.
- [23] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, “A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries,” *J Am Med Inform Assoc*, vol. 18, no. 5, pp. 601–606, Oct. 2011.
- [24] B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu, “Clinical entity recognition using structural support vector machines with rich features,” in *Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics*, New York, NY, USA, 2012, pp. 13–20.
- [25] B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu, “Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features,” *BMC Medical Informatics and Decision Making*, vol. 13, no. Suppl 1, p. S1, Apr. 2013.
- [26] B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, and H. Xu, “A hybrid system for temporal information extraction from clinical text,” *J Am Med Inform Assoc*, Apr. 2013.