Information **2015**, *6*, 848-865; doi:10.3390/info6040848

OPEN ACCESS

information

ISSN 2078-2489

www.mdpi.com/journal/information

Article

Effects of Semantic Features on Machine Learning-Based Drug Name Recognition Systems: Word Embeddings *vs.* Manually Constructed Dictionaries

Shengyu Liu, Buzhou Tang, Qingcai Chen * and Xiaolong Wang

Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China; E-Mails: shengyu_liu@163.com (S.L.); tangbuzhou@gmail.com (B.T.); wangxl@insun.hit.edu.cn (X.W.)

* Author to whom correspondence should be addressed; E-Mail: qingcai.chen@gmail.com or qingcai.chen@hitsz.edu.cn; Tel.: +86-755-2603-3475; Fax: +86-755-2603-3182.

Academic Editors: Yong Yu and Yu Wang

Received: 17 October 2015 / Accepted: 4 December 2015 / Published: 11 December 2015

Abstract: Semantic features are very important for machine learning-based drug name recognition (DNR) systems. The semantic features used in most DNR systems are based on drug dictionaries manually constructed by experts. Building large-scale drug dictionaries is a time-consuming task and adding new drugs to existing drug dictionaries immediately after they are developed is also a challenge. In recent years, word embeddings that contain rich latent semantic information of words have been widely used to improve the performance of various natural language processing tasks. However, they have not been used in DNR systems. Compared to the semantic features based on drug dictionaries, the advantage of word embeddings lies in that learning them is unsupervised. In this paper, we investigate the effect of semantic features based on word embeddings on DNR and compare them with semantic features based on three drug dictionaries. We propose a conditional random fields (CRF)-based system for DNR. The skip-gram model, an unsupervised algorithm, is used to induce word embeddings on about 17.3 GigaByte (GB) unlabeled biomedical texts collected from MEDLINE (National Library of Medicine, Bethesda, MD, USA). The system is evaluated on the drug-drug interaction extraction (DDIExtraction) 2013 corpus. Experimental results show that word embeddings significantly improve the performance of the DNR system and they are competitive with semantic features based on drug dictionaries. F-score is improved by 2.92 percentage points when word embeddings are added into the baseline system. It is comparative with the improvements from semantic features based on drug dictionaries. Furthermore, word embeddings are complementary to the semantic features based on drug dictionaries. When both word embeddings and semantic features based on drug dictionaries are added, the system achieves the best performance with an F-score of 78.37%, which outperforms the best system of the DDIExtraction 2013 challenge by 6.87 percentage points.

Keywords: drug name recognition; word embeddings; drug information extraction; biomedical texts

1. Introduction

Drug name recognition (DNR) is a critical step for drug information extraction such as drug interactions [1]. It contains two tasks: detecting the boundaries of drug names in unstructured texts (drug detection) and classifying the detected drug names into some predefined categories (drug classification). It is a challenging task for various reasons. First, new drugs are constantly and rapidly discovered. Second, naming conventions are not strictly followed, although they are available for a variety of domains in the biomedicine field.

There has been a lot of research on DNR, including the drug-drug interaction extraction (DDIExtraction) 2013 challenge [2]. A variety of methods proposed for DNR mainly fall into three categories: (i) dictionary-based methods that utilize lists of terms from different resources to identify drug names in biomedical texts [3]; (ii) ontology-based methods that map each unit of a text into one or more domain-specific concepts, and then combine the concepts into drug names by manually defined rules [1]; and (iii) machine learning-based methods that build machine learning models based on labeled corpora to identify drug names [4]. Among them, machine learning-based methods are superior to the other two categories of methods because of their good performances and robustness when a large labeled corpus is available [2,5].

DNR is a typical named entity recognition (NER) task and is mostly regarded as a sequence labeling problem. Lots of machine learning algorithms have been used for DNR such as support vector machines (SVM) [6] and conditional random fields (CRF) [4]. CRF is the top choice for DNR since it is one of the most reliable sequence labeling algorithms and has shown good performances on a large number of NER tasks such as NER in the newswire domain [7,8], biomedical NER [9,10] and clinical NER [11,12]. Besides machine learning algorithms, features are another key factor for DNR. The semantic features based on drug dictionaries are very important, but it is not easy to get them. Firstly, it requires a large number of domain experts to spend an enormous amount of time to build large-scale drug dictionaries. Secondly, adding new developed drugs to existing drug dictionaries is necessary. However, it is difficult to update the dictionaries immediately after new drugs are developed. Therefore, it is certainly worth seeking new methods that are not only able to generate semantic features automatically from raw or unlabeled biomedical texts, but also able to update the semantic features when new biomedical texts are generated.

Word embeddings induced by unsupervised deep learning algorithms on large-scale unlabeled samples have been widely used in various natural language processing tasks recently and shown stable improvements [13–16]. They contain rich latent semantic information of words and have potential of improving the performances of machine learning-based DNR systems. It is worth investigating the effects of unsupervised semantic features based on word embeddings on machine learning-based DNR systems, especially, when they are compared with semantic features based on manually constructed drug dictionaries: whether word embeddings can replace the semantic features based on drug dictionaries. However, this has not been investigated.

In this paper, we investigate the effect of word embeddings on machine learning-based DNR systems, and compare them with the semantic features from different drug dictionaries. The machine learning algorithm used in our systems is CRF, the word embeddings are induced by the skip-gram model [17,18] on about 17.3 GB unlabeled biomedical texts collected from MEDLINE, and three common drug dictionaries (*i.e.*, a drug dictionary from United States (U.S.) Food and Drug Administration (FDA), DrugBank [19] and Jochem [20]) are used to extract other semantic features. Experimental results on the DDIExtraction 2013 corpus show that word embeddings significantly improve the performance of the CRF-based DNR system. F-score is improved by 2.92 percentage points when word embeddings are added into the baseline system that does not use any semantic features. It is comparative with the improvements from semantic features based on the three drug dictionaries. Furthermore, word embeddings are complementary to the semantic features based on the drug dictionaries. When both types of semantic features are added to the baseline system at the same time, the system achieves the best performance with an F-score of 78.37%, which outperforms the best system of the DDIExtraction 2013 challenge by 6.87 percentage points.

2. Related Work

2.1. Drug Name Recognition

The early methods for DNR are mainly dictionary-based or ontology-based because of lacking manually annotated corpora. Segura-Bedmar *et al.* [1] presented a two-layer system for DNR in biomedical texts. The first layer is the Unified Medical Language System (UMLS) MetaMap Transfer (MMTx) program [21] that maps biomedical texts to UMLS concepts. The subsequent layer defines nomenclature rules recommended by the World Health Organization (WHO) International Nonproprietary Names (INNs) Program to filter drugs from all concepts and adjust their classes.

To accelerate the development of related researches on drugs, MAVIR research network and University Carlos III of Madrid in Spain launched two challenges in 2011 and 2013, respectively, *i.e.*, DDIExtraction 2011 and DDIExtraction 2013. Both challenges provided manually annotated corpora that contain drug names. The DDIExtraction 2011 challenge [22] was designed to extract drug-drug interactions from biomedical texts. Only the mentions of drugs without class information are labeled. Based on this corpus, He *et al.* [4] presented a machine learning-based method for drug detection. They built a drug dictionary using a semi-supervised learning method and integrated it into a CRF-based system. The DDIExtraction 2013 challenge also focused on extraction of drug–drug interactions, but DNR was proposed as an individual subtask, where both mentions and classes of drugs are labeled. Four classes of drugs are defined, namely, "*drug*", "*brand*", "*group*" and "*no-human*". Six teams participated in the DNR subtask of the DDIExtraction 2013 challenge. Two classes of systems are presented for the DNR subtask in this challenge: dictionary-based and machine learning-based systems. The best system is based on CRF. Features used in the proposed systems include word, part-of-speech (POS), affix feature, orthographical feature, word shape feature and dictionary feature, *etc.* However, word embeddings that contain rich latent semantic information were not used for DNR.

To promote the development of systems for recognition of chemical entities, the Critical Assessment of Information Extraction systems in Biology (BioCreAtIvE) IV organized the chemical compound and drug name recognition (CHEMDNER) task [5]. The CHEMDNER task was divided into two subtasks: chemical document indexing and chemical entity mention recognition. Many systems were proposed for the chemical entity mention recognition subtask and the top-ranked systems are also based on machine learning algorithms [23–27]. However, the chemical entity mention recognition subtask only focused on detecting boundaries of chemicals and drugs in texts and did not request predictions of the classes of them. It can be regarded as a drug detection task. The DNR task in this study is different from the chemical entity mention recognition subtask, because it focuses on detecting and classifying drugs simultaneously. Moreover, the DNR task is much more difficult than the chemical entity mention recognition task. For example, the best performing system for the chemical entity mention recognition subtask achieved an F-score of 87.39% [23]. The best performing system for the DDIExtraction 2013 challenge achieved an F-score of 83.30% when only considering boundaries of the detected drugs. However, the F-score dropped to 71.50% when both boundaries and classes were considered. Word embeddings were used in some systems for the chemical entity mention recognition subtask [24,25]. However, the effects of word embeddings on the drug detection task were limited. For example, in [24], F-score is improved by 0.19 percentage points (from 84.96% to 85.15%) when word embeddings are added. In this study, word embeddings are used for the DNR task, which detects and classifies drugs simultaneously. Experimental results of this study demonstrate that word embeddings are highly beneficial to the DNR task.

The studies described above focus on drug name recognition in biomedical literature. There are also studies that extract drug names from clinical text such as discharge summaries [11,28–30] and clinical notes [31,32]. Many different methods have been proposed for drug name recognition in clinical text. For example, Xu *et al.* [28] proposed a dictionary-based method for drug name recognition in discharge summaries. Drug names are recognized by matching a drug dictionary against the discharge summaries. Patrick *et al.* [11] used a CRF-based method to extract drug names from discharge summaries. Doan *et al.* [29] combined a dictionary-based method, a SVM-based method and a CRF-based method into a voting system to identify drug names. The voting system could achieve better performance than each single method.

2.2. Word Embeddings Learning Algorithms

Many methods have been proposed to induce unsupervised word representations. Word representations can be classified into three categories: clustering-based word representations, distributional word representations and distributed word representations [13]. Clustering-based word representations such as Brown clustering [33] induce clusters over words. Each word is represented by the clusters it belongs to. Distributional word representations such as LSA [34], HAL [35] and random indexing [36] reduce a high-dimensional word co-occurrence matrix to a low-dimensional semantic matrix, in which each word corresponds to a low-dimensional vector. Distributed word representations (*i.e.*, word embeddings) generate a low-dimensional, real-valued and dense vector for each word using neural language models. The vectors can capture rich latent semantic information of words. Bengio *et al.* [37] proposed a neural network architecture to predict the next word given the previous ones. Collobert *et al.* [14] proposed another neural network architecture that checks whether a text fragment is valid. The subsequent studies were mainly based on these two architectures. To speed up Bengio and coworker's architecture, a hierarchical neural language model was presented in [38,39]. Huang *et al.* [40] added global text into Collobert and coworker's architecture to improve word embeddings. Mikolov *et al.* [41] proposed a recurrent neural network language model based on Bengio and coworker's architecture, where the size of context is not limited and the word itself is also a part of its context. The main problem of these word embeddings learning algorithms is that they are too time-consuming. Recently, Mikolov *et al.* [17,18] simplified Bengio and coworker's architecture, and presented two novel models: continuous bag-of-words (CBOW) model and skip-gram model with much lower computational cost. The word embeddings induced by these two models are also of high quality as previous models. As reported in [17], the skip-gram model is slightly superior to the CBOW model. Due to the effectiveness and efficiency of the skip-gram model, we use it to induce word embeddings on 17.3 GB article abstracts extracted from MEDLINE in our study.

3. Methods

Figure 1 shows the architecture of our CRF-based DNR system, which consists of four components: (1) a preprocessing module to split sentences and tokenize them using the NLTK toolkit [42]; (2) a feature extraction module to extract features including word, POS, affix feature, orthographical feature, word shape feature and dictionary feature, *etc.*; (3) a CRF classifier for DNR; (4) a postprocessing module to generate drug names from labeled words. The CRF classifier and features are described below in detail.

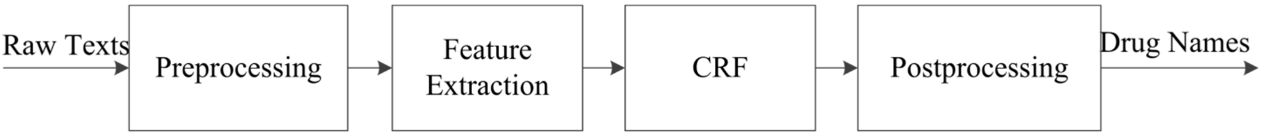

Figure 1. Architecture of our conditional random fields (CRF)-based DNR system.

3.1. Drug Name Recognition

DNR is usually treated as a sequence labeling problem. Tokens in a sentence are labeled with different tags. Each tag contains information about whether a token is part of a drug name and its position in a drug name. Several tagging schemes proposed for other NER tasks [43] are also suitable for DNR. In this paper, we adopt **BILOU**, a representative tagging scheme, to represent drugs. In this scheme, **BILOU**, respectively, represent the beginning (**B**) of an entity, inside (**I**) of an entity, last token of an entity (**L**), outside of an entity (**O**) and a single-token entity (**U**). In the DDIExtraction 2013 challenge, drugs are divided into four types: "*drug*", "*brand*", "*group*" and "*no-human*". Therefore, 17 tags are needed to represent all drugs. The tags are: B-drug, I-drug, L-drug, U-drug, B-brand, I-brand, L-brand, U-brand, B-group, I-group, L-group, U-group, B-no-human, I-no-human, L-no-human, U-no-human and O. Figure 2 shows an example of a sentence labeled by **BILOU**.

Sentence: The risk of myopathy is increased with concurrent administration of cyclosporine, fibric acid derivatives, azole antifungals.

BILOU: The\O risk\O of\O myopathy\O is\O increased\O with\O concurrent\O administration\O of\O cyclosporine\U-drug ,\O fibric\B-group acid\I-group derivatives\L-group ,\O azole\B-group antifungals\L-group .\O

Figure 2. An example of a sentence labeled by **BILOU**.

3.2. Conditional Random Fields

CRF is a typical sequence labeling algorithm [44]. It has been widely applied to a large number of Natural Language Processing (NLP) tasks such as named entity recognition [45], shallow parsing [46], and Chinese word segmentation [47] and has shown good performances on these tasks. The task of sequence labeling problem is to assign a sequence of labels **y** = {y_1, y_2, …, y_n} to a sequence of input **x** = {x_1, x_2, …, x_n}. In the case of DNR, **x** corresponds to a tokenized sentence, while **y** corresponds to a sequence of tags mentioned in the previous section. CRF adopts a conditional probability distribution to model the sequence labeling problem as follows

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) \tag{1}$$

In first-order linear-chain CRF, the conditional probability $p(\mathbf{y}|\mathbf{x})$ is defined as

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{i=1}^{n} \exp\left(\sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, \mathbf{x}, i) + \sum_{k=1}^{l} \mu_k f_k(y_i, \mathbf{x}, i)\right) \tag{2}$$

where Z(**x**) is an input-dependent normalization factor, f_j is a transition feature function, which captures the properties of the input tokens and the labels of tokens at position i-1 and i, f_k is a state feature function, which captures the properties of the input tokens and the label of token at position i, m and l are the numbers of transition feature function and state feature function. λ_j and μ_k are the weights of f_j and f_k, which are learned from the training data.

In our study, CRFsuite [48] is used as an implementation of first-order linear-chain CRF.

3.3. Skip-Gram Model

Mikolov *et al.* [17,18] proposed two efficient word embeddings learning models: CBOW and skip-gram. The idea of the CBOW model is to predict a word based on its context. Contrary to the CBOW model, the skip-gram model predicts a word's context based on the word itself. Since better performance for skip-gram model is reported in [17], we use the skip-gram model to induce word embeddings in our experiments.

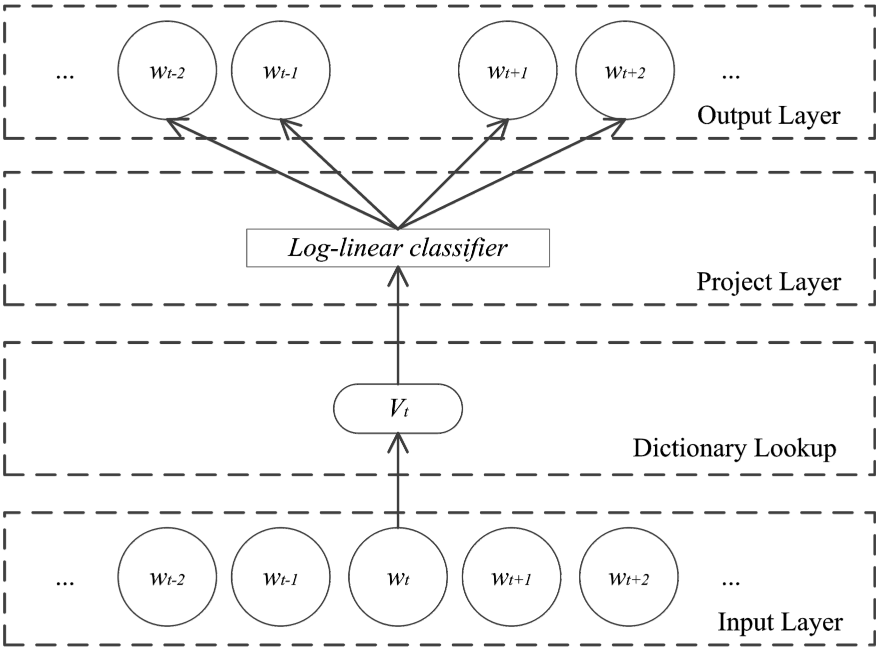

Figure 3. Architecture of the skip-gram model.

Figure 3 shows the architecture of the skip-gram model. The circles denote the literal representations of words and the ovals denote the vector representations of words. Any word w_t with embeddings V_t is input into a log-linear classifier to predict its context, *i.e.*, w_{t-2}, w_{t-1}, w_{t+1} and w_{t+2} when the words in the [−2, 2] window are considered. The objective function of the skip-gram model is

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0}\log p(w_{t+j}\mid w_t) \tag{3}$$

where T is the size of word sequence in the training corpus and c is the length of context for word w_t.

In the skip-gram model, $p(w_{t+j}|w_t)$ is defined using the softmax function:

$$p(w_{t+j}\mid w_t)=\frac{\exp(v_{w_{t+j}}^{'\ T}v_{w_t})}{\sum_{w=1}^{N}\exp(v_w^{'\ T}v_{w_t})} \tag{4}$$

where v_w and $v_w^{'}$ are the vector representations of w when it is at the input and output layers respectively, and N is the size of word vocabulary.

In our study, word2vec tool [49] is used as an implementation of the skip-gram model.

3.4. Features

The features used in this study include general features that are commonly used for NER tasks, semantic features based on three existing drug dictionaries, and semantic features based on word embeddings induced by the skip-gram model. All of them are described below in detail.

3.4.1. General Features

The general features used in this study are as follows:

1) Word feature: The word itself.

2) POS feature: The POS type of a word generated by the GENIA toolkit [50], where 36 POS types and 12 other types (for punctuation and currency symbols) defined in Penn Treebank [51] tagset are used.
3) Chunk feature: The chunk information of a word generated by the GENIA toolkit, where 11 chunking types defined in the CoNLL-2000 shared task [52] are used.
4) Affix feature: Prefixes and suffixes of the length of 3, 4, and 5.
5) Orthographical feature: Each word is classified into one of the four classes {"Is-capitalized", "All-capitalized", "Alphanumeric", "All-digits"} based on regular expressions. The class label of a word is used as its feature. In addition, we check whether a word contains a hyphen, and denote whether a word contains a hyphen or not by {"Y", "N"}.
6) Word shape feature: We used two types of word shapes: "generalized word class" and "brief word class" similar to [9]. The generalized word class of a token is generated by mapping any uppercase letter, lowercase letter, digit and other character in this token to "X", "x", "0" and "O", respectively, while the brief word class of a token is generated by mapping consecutive uppercase letters, lowercase letters, digits and other characters to "X", "x", "0" and "O", respectively. For example, word shape features of "Aprepitant" are "Xxxxxxxxxx" and "Xx", respectively.

3.4.2. Semantic Features Based on Drug Dictionaries

In this paper, we compare the semantic features based on three drug dictionaries with semantic features based on word embeddings. The semantic features based on drug dictionaries are generated in the way similar to [6] that denotes whether a word appears in a drug dictionary by {"Y", "N"}. The detailed descriptions of each of the three drug dictionaries are as follows:

1) DrugBank: DrugBank [19] is a structured bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information [53]. It contains 6825 drug entries and all drug names in it are extracted to build a drug dictionary.
2) Drugs@FDA: Drugs@FDA [54] is a database provided by U.S. FDA. It contains information about FDA-approved drug names, generic prescription, *etc.* We extract the *Drugname* and *Activeingred* fields in Drugs@FDA and totally 8391 drug names are extracted to build a drug dictionary.
3) Jochem: Jochem [20] is a joint chemical dictionary for the identification of small molecules and drugs in text. We extract 1,527,751 concepts from Jochem.

3.4.3. Semantic Features Based on Word Embeddings

The semantic features based on word embeddings induced by the skip-gram model are as follows:

Rather than directly using the word embeddings, we group words in the vocabulary into different semantic classes according to their word embeddings, and use the semantic class that a word belongs to as its latent semantic feature. The word embeddings are induced on the MEDLINE abstracts using the skip-gram model. MEDLINE is the U.S. National Library of Medicine journal citation database that provides over 21 million references to biomedical and life sciences journal articles back to 1946. It is a rich source of biomedical texts that contains citations from over 5600 scholarly journals published around the world. To get unlabeled biomedical texts, we download the MEDLINE released in 2013 [55] and extract all article abstracts. Finally, in total, we obtain 17.3 GB unlabeled biomedical texts. All of them are tokenized using the NLTK toolkit. The tokenized texts contain 110 million sentences with 2.8 billion words. The size of word vocabulary is 10.8 million. Following previous studies [15,56], the dimension of word embeddings is set to 50. Then we run the skip-gram model on the MEDLINE corpus. The whole training process only takes about one hour. To group words into semantic classes, we adopt k-means clustering algorithm.

4. Experiments

To investigate the effect of semantic features based on word embeddings on DNR, we start with the baseline system that uses general features described in Section 3.4.1, and then gradually add semantic features based on each drug dictionary and the semantic features based on word embeddings. All experiments are conducted on the DDIExtraction 2013 corpus. Each CRF model uses default parameters except the L2 regularization coefficient. The optimal L2 regularization coefficient is selected from {0.1, 0.2, …, 1.0} via 10-fold cross validation on the training set of the DDIExtraction 2013 challenge. In the case of k-means algorithm, the optimal number of clusters is selected from {100, 200, 300, ..., 1000} via 10-fold cross validation on the training set of the DDIExtraction 2013 challenge. It is 400. We firstly determine the optimal number of clusters in k-means clustering algorithm with the default L2 regularization coefficient in CRFsuite. Then, we determine the optimal L2 regularization coefficient using the optimal number of clusters.

4.1. Dataset

The DDIExtraction 2013 corpus composes of 826 documents with 7641 sentences. The documents come from two sources: DrugBank and MEDLINE. Four types of drugs are manually annotated, and the number of drugs is 15,450. The corpus contains a training set and a test set. The training set and test set are used for system development and system evaluation, respectively. Both of them contain documents from DrugBank and MEDLINE. The definitions of the four types of drugs are presented in detail as follows:

- Drug: Names of chemical agents used in the treatment, cure, prevention or diagnosis of diseases that have been approved for human use.
- Brand: Brand names of drugs specified by pharmaceutical companies.
- Group: Terms in texts designating chemical or pharmacological relationships among a group of drugs.
- No-human: Names of chemical agents that affect living organism, but have not been approved to be used in humans for a medical purpose.

Table 1 shows the statistics of corpus of the DDIExtraction 2013 challenge.

Table 1. Statistics of the drug-drug interaction extraction (DDIExtraction) 2013 corpus.

	DrugBank			**MEDLINE**		
	Training	**Test**	**Total**	**Training**	**Test**	**Total**
Documents	572	54	626	142	58	200
Sentences	5675	145	5820	1301	520	1821
Drug	8197	180	8377	1228	171	1399
Group	3206	65	3271	193	90	283
Brand	1423	53	1476	14	6	20
No-human	103	5	108	401	115	516

4.2. Evaluation Metrics

We use precision (P), recall (R) and F-score (F1) to evaluate the performance of a DNR system under the four evaluation criteria provided by the DDIExtraction 2013 challenge. These criteria are:

- Strict matching: A tagged drug name is correct only if its boundary and class exactly match with a gold drug name.
- Exact boundary matching: A tagged drug name is correct if its boundary matches with a gold drug name regardless of its class.
- Type matching: A tagged drug name is correct if there is some overlap between it and a gold drug name of the same class.
- Partial boundary matching: A tagged drug name is correct if there is some overlap between it and a gold drug name regardless of its class.

In the DDIExtraction 2013 challenge, the strict matching criterion is the primal one for the DNR task.

4.3. Experimental Results

We first compare the performances of the CRF-based DNR systems when different semantic features, including semantic features based on each one of the three drug dictionaries and word embeddings, are added into the baseline system.

Table 2 shows the experimental results under the strict matching criterion, where F_g denotes the features used in the baseline system including word, POS, chunk features, affix features, orthographical features and word shapes. F_{FDA}, $F_{DrugBank}$ and F_{Jochem} denote the semantic features based on Drugs@FDA, DrugBank and Jochem, respectively. F_{WE} denotes the semantic features based on word embeddings.

It can be seen that both the semantic features based on drug dictionaries and the semantic features based on word embeddings are beneficial to DNR. When F_{FDA}, $F_{DrugBank}$ and F_{Jochem} are added to F_g, F1 is improved by 2.87 percentage points (from 72.71% to 75.58%), 3.54 percentage points (from 72.71% to 76.25%) and 2.12 percentage points (from 72.71% to 74.83%), respectively. When F_{WE} is added to F_g, F1 is improved by 2.92 percentage points (from 72.71% to 75.63%), which is comparative to the improvements coming with F_{FDA}, $F_{DrugBank}$ and F_{Jochem}, respectively. The differences between F-scores of systems using and not using semantic features are significant (p-value < 0.05) according to approximate randomization [29].

When the semantic features based on drug dictionaries and word embeddings are added into the baseline system at the same time, the performance of the DNR system is further improved. For example, when both $F_{DrugBank}$ and F_{WE} are added ($F_g + F_{DrugBank} + F_{WE}$), the F1 achieves 78.05%, which is higher than that when only $F_{DrugBank}$ or F_{WE} is added (76.25% and 75.63%). When all sematic features are added ($F_g + F_{FDA} + F_{DrugBank} + F_{Jochem} + F_{WE}$), the CRF-based DNR system achieves the highest F1 of 78.37%, which is higher than that of the system when all semantic features based on the three drug dictionaries are added ($F_g + F_{FDA} + F_{DrugBank} + F_{Jochem}$) (77.39%).

Another interesting finding is that the system using all semantic features based on the three drug dictionaries outperforms the systems using the semantic features based on only one drug dictionary. The differences of F1 range from 1.14% to 2.56%.

Table 2. Performances of the CRF-based DNR systems under the strict matching criterion when different features are used (%).

Feature	P	R	F1
F_g	78.41	67.78	72.71
$F_g + F_{WE}$	82.70	69.68	75.63
$F_g + F_{FDA}$	83.19	69.24	75.58
$F_g + F_{DrugBank}$	85.51	68.80	76.25
$F_g + F_{Jochem}$	77.71	72.16	74.83
$F_g + F_{FDA} + F_{WE}$	85.59	70.12	77.09
$F_g + F_{DrugBank} + F_{WE}$	86.24	71.08	78.05
$F_g + F_{Jochem} + F_{WE}$	79.64	71.28	75.23
$F_g + F_{FDA} + F_{DrugBank} + F_{Jochem}$	83.84	71.87	77.39
$F_g + F_{FDA} + F_{DrugBank} + F_{Jochem} + F_{WE}$	84.75	72.89	78.37

4.4. Performance Comparison between Our System and Participating Systems of DDIExtraction 2013

To further investigate our system, we compare our best system with all systems in the DDIExtraction 2013 challenge. Table 3 shows the comparison results under the strict matching criterion. It is shown that our system outperforms WBI that ranked first in the DDIExtraction 2013 challenge by 6.87 percentage points in F1. The detailed comparisons between them are shown in Table 4, where the overall performance under the four criteria, the overall performance under the strict matching criterion on the test subsets from DrugBank and MEDLINE and the performance of each type under the strict matching criterion are listed. The last six lines are the P, R and F1 under the strict matching criterion. Our system outperforms WBI, on the whole, when the classes of drugs are considered. The differences of F1 under the strict matching and type matching criteria are 6.87% and 6.08%, respectively. If the classes of drugs are not considered, the differences of F1 are very small. The differences of F1 under the exact matching and partial matching criteria are only 0.56% and 0.18%. For each type of drugs, our system achieves much better performance. The differences of F1 range from 7.93% (for Group) to 13.79% (for Brand). On both DrugBank and MEDLINE test subsets, our system also outperforms WBI by 1.9 percentage points and 10.15 percentage points in F1, respectively.

Table 3. Performance comparison between our system and systems of DDIExtraction 2013 (%).

System	Strict		
	P	**R**	**F1**
Our system	84.75	72.89	78.37
WBI [57]	73.40	69.80	71.50
NLM_LHC	73.20	67.90	70.40
LASIGE [58]	69.60	62.10	65.60
UTurku [6]	73.70	57.90	64.80
UC3M [59]	51.70	54.20	52.90
UMCC_DLSI_DDI [60]	19.50	46.50	27.50

Table 4. Detailed comparison between our system and the best system of DDIExtraction 2013 (%).

	WBI			Our System			
	P	**R**	**F1**	**P**	**R**	**F1**	**ΔF1**
Strict	73.40	69.80	71.50	84.75	72.89	78.37	+6.87
Exact	85.50	81.30	83.30	90.68	77.99	83.86	+0.56
Type	76.70	73.00	74.80	87.46	75.22	80.88	+6.08
Partial	87.70	83.50	85.60	92.37	79.45	85.42	−0.18
Drug (strict)	73.60	85.20	79.00	92.02	85.47	88.62	+9.62
Brand (strict)	81.00	86.40	83.60	100.00	94.92	97.39	+13.79
Group (strict)	79.20	76.10	77.60	89.44	81.94	85.53	+7.93
No-human (strict)	31.40	9.10	14.10	89.47	14.05	24.29	+10.19
DrugBank (strict)	88.10	87.50	87.80	90.60	88.82	89.70	+1.90
MEDLINE (strict)	60.70	55.80	58.10	78.77	60.21	68.25	+10.15

5. Discussion

In this paper, we propose a CRF-based system for DNR, investigate the effect of semantic features based on word embeddings on DNR, and compare the semantic features based on word embeddings with semantic features based on three drug dictionaries. To our best knowledge, although some distributional word representations have been used for DNR [31] and word embeddings have been widely used in other NLP tasks, it is the first time to investigate the effects of word embeddings on DNR. The word embeddings used in our experiments are induced by the skip-gram model, a novel word embeddings learning algorithm proposed recently, on about 17.3 GB unlabeled biomedical texts collected from MEDLINE. Experimental results on the corpus of the DDIExtraction 2013 challenge show that the semantic features based on word embeddings are beneficial to DNR. The improvement from word embeddings achieves 2.92 percentage points when they are added into the baseline system. It is competitive with the semantic features based on three drug dictionaries: Drugs@FDA, DrugBank and Jochem. Furthermore, when both types of semantic features are added to the baseline system at the same time, the performance is further improved. The highest F-score of our system achieves 78.37%, better than that of the best system of the DDIExtraction 2013 challenge by 6.87 percentage points.

In [13], two different word embeddings are integrated into a CRF-based system for NER on the CoNLL-2003 shared task dataset drawn from the Reuters newswire [61]. F-score of the system is improved by 1.61 and 1.44 percentage points when the two word embeddings are added into the system, respectively. It can be seen that the improvement obtained with word embeddings for DNR is comparable with that for NER in the newswire domain.

It is easy to understand that word embeddings improve the performance of the CRF-based DNR system since they capture rich latent semantic information to make sure similar words close with each other. Especially for the rare words in the training and test sets, the word embeddings induced on a large-scale unlabeled corpus naturally perform smoothing as they are dense, real-valued vectors. For this reason, when the semantic features based on word embeddings are added, the recall is improved. For example, the recall is improved by 1.90 percentage points when the semantic features based on word embeddings are added into the baseline system as shown in Table 2.

Compared to the semantic features based on a single drug dictionary, the semantic features based on word embeddings show competitive effects. It indicates that word embeddings can work as a drug dictionary in machine learning-based DNR systems. In fact, it is too difficult to manually build a complete drug dictionary that covers all other drug dictionaries. In the case of the three drug dictionaries used in our system, they are complementary to each other as the system using the semantic features based on all of them shows better performance than using the semantic features based on one of them (“$F_g + F_{FDA} + F_{DrugBank} + F_{Jochem}$” *vs.* {“$F_g + F_{FDA}$”, “$F_g + F_{DrugBank}$”, “$F_g + F_{Jochem}$”} in Table 2). Similarly, the semantic features based on word embeddings are also complementary to the semantic features based on drug dictionaries.

Compared to the best system in the DDIExtraction 2013 challenge (*i.e.*, WBI), our best system achieves much better performance. However, it is not good enough for medical information extraction systems and applications, e.g., drug–drug interaction extraction, clinical decision support and drug development. Firstly, the performance for “*no-human*” is still very bad though it is much better than WBI. The primary reason for this is that four types of drugs in the corpus are extremely imbalanced. “*no-human*” accounts for about 4% (624/15,450), while “*drug*” accounts for about 63% (9776/15,450). We will study this data imbalance problem for improvement in the future. Secondly, the performance on the test subset from MEDLINE is much lower than that on the test subset from DrugBank (68.25% *vs.* 89.70%). In part, that is because sentences in DrugBank are much different from those in MEDLINE. When the annotated sentences in MEDLINE are not enough, a large number of annotated sentences in DrugBank are of limited help to improve the performance on MEDLINE. The most direct approach is to annotate more sentences in MEDLINE.

In our study, only the skip-gram model is adopted to induce word embeddings. It is worth comprehensively comparing different word embeddings learning algorithms on DNR, which is another case of our future work.

6. Conclusions

In this paper, we propose a CRF-based DNR system using semantic features based on word embeddings induced by the skip-gram model and compare the effect of word embeddings with that of the semantic features based on drug dictionaries on the DNR system. Experiments on the DDIExtraction 2013 corpus show that semantic features based on word embeddings are beneficial to DNR and complementary to semantic features based on drug dictionaries. For future work, it is worth investigating other word embeddings learning algorithms on DNR.

Acknowledgments

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

Author Contributions

The work presented here was a collaboration of all the authors. All authors contributed to designing the methods and experiments. Shengyu Liu performed the experiments. Shengyu Liu, Buzhou Tang, Qincai Chen and Xiaolong Wang analyzed the data and interpreted the results. Shengyu Liu wrote the paper. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Segura-Bedmar, I.; Martínez, P.; Segura-Bedmar, M. Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems. *Drug Discov. Today* **2008**, *13*, 816–823.
2. Segura-Bedmar, I.; Martínez, P.; Herrero-Zazo, M. SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 341–350.
3. Sanchez-Cisneros, D.; Martínez, P.; Segura-Bedmar, I. Combining dictionaries and ontologies for drug name recognition in biomedical texts. In Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics, San Francisco, CA, USA, 1 November 2013; pp. 27–30.
4. He, L.; Yang, Z.; Lin, H.; Li, Y. Drug name recognition in biomedical texts: A machine-learning-based method. *Drug Discov. Today* **2014**, *19*, 610–617.
5. Krallinger, M.; Leitner, F.; Rabal, O.; Vazquez, M.; Oyarzabal, J.; Valencia, A. CHEMDNER: The drugs and chemical names extraction challenge. *J. Cheminformatics* **2015**, *7* (Suppl. 1), S1.
6. Björe, J.; Kaewphan, S.; Salakoski, T. UTurku: Drug named entity detection and drug-drug interaction extraction using SVM classification and domain knowledge. In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 651–659.
7. Finkel, J.; Grenager, T.; Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, USA, 25–30 June 2005; pp. 363–370.
8. Tkachenko, M.; Simanovsky, A. Named entity recognition: Exploring features. In Proceedings of the KONVENS 2012, Vienna, Austria, 19–21 September 2012; pp. 118–127.

9. Settles, B. Biomedical named entity recognition using conditional random fields and rich feature sets. In Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland, 23–27 August 2004; pp. 104–107.
10. McDonald, R.; Pereira, F. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinform.* **2005**, *6* (Suppl. 1), S6.
11. Patrick, J.; Li, M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 524–527.
12. Jiang, M.; Chen, Y.; Liu, M.; Rosenbloom, S.T.; Mani, S.; Denny, J.C.; Xu, H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 601–606.
13. Turian, J.; Ratinov, L.; Bengio, Y. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010; pp. 384–394.
14. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
15. Tang, B.; Cao, H.; Wang, X.; Chen, Q.; Xu, H. Evaluating word representation features in biomedical named entity recognition tasks. *Biomed. Res. Int.* **2014**, *2014*, doi:10.1155/2014/240403.
16. Passos, A.; Kumar, V.; McCallum, A. Lexicon Infused Phrase Embeddings for Named Entity Resolution. In Proceedings of the 18th Conference on Computational Language Learning, Baltimore, MD, USA, 26–27 June 2014; pp. 78–86.
17. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the Workshop at ICLR, Scottsdale, AZ, USA, 2–4 May 2013.
18. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, CA, USA, 5–10 December 2013; pp. 3111–3119.
19. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; *et al.* DrugBank 3.0: A comprehensive resource for "omics" research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035–D1041.
20. Hettne, K.; Stierum, R.; Schuemie, M.; Hendriksen, P.; Schijvenaars, B.; van Mulligen, E.; Kleinjans, J.; Kors, J. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **2009**, *25*, 2983–2991.
21. Aronson, A.; Bodenreider, O.; Chang, H.; Humphrey, S.; Mork, J.; Nelson, S.; Rindflesch, T.; Wilbur, W. The NLM indexing initiative. In Proceedings of the AMIA Annual Symposium, Los Angeles, CA, USA, 4–8 November 2000; pp. 17–21.
22. Segura-Bedmar, I.; Martínez, P.; Sánchez-Cisneros, D. The 1st DDIExtraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts. In Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction, Huelva, Spain, 7 September 2011; pp. 1–9.
23. Leaman, R.; Wei, C.; Lu, Z. tmChem: A high performance approach for chemical named entity recognition and normalization. *J. Cheminformatics* **2015**, *7* (Suppl. 1), S3.

24. Tang, B.; Feng, Y.; Wang, X.; Wu, Y.; Zhang, Y.; Jiang, M.; Wang, J. A comparison of conditional random fields and structured support vector machines for chemical entity recognition in biomedical literature. *J. Cheminformatics* **2015**, *7* (Suppl. 1), S8.
25. Lu, Y.; Ji, D.; Yao, X.; Wei, X.; Liang, X. CHEMDNER system with mixed conditional random fields and multi-scale word clustering. *J. Cheminformatics* **2015**, *7* (Suppl. 1), S4.
26. Batista-Navarro, R.; Rak, R.; Ananiadou, S. Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. *J. Cheminformatics* **2015**, *7* (Suppl. 1), S6.
27. Campos, D.; Matos, S.; Oliveira, J. A document processing pipeline for annotating chemical entities in scientific documents. *J. Cheminformatics* **2015**, *7* (Suppl. 1), S7.
28. Xu, H.; Stenner, S.; Doan, S.; Johnson, K.B.; Waitman, L.R.; Denny, J.C. MedEx: A medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 19–24.
29. Doan, S.; Collier, N.; Xu, H.; Duy, P.; Phuong, T. Recognition of medication information from discharge summaries using ensembles of classifiers. *BMC Med. Inform. Decis. Mak.* **2012**, *12*, doi:10.1186/1472-6947-12-36.
30. Halgrim, S.; Xia, F.; Solti, I.; Cadag, E.; Uzuner, Ö. A cascade of classifiers for extracting medication information from discharge summaries. *J. Biomed. Semant.* **2011**, *2* (Suppl. 3), S2.
31. Henriksson, A.; Kvist, M.; Dalianis, H.; Duneld, M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *J. Biomed. Inform.* **2015**, *57*, 333–349.
32. Skeppstedt, M.; Kvist, M.; Nilsson, G.; Dalianis, H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *J. Biomed. Inform.* **2014**, *49*, 148–158.
33. Brown, P.; de Souza, P.; Mercer, R.; Pietra, V.; Lai, J. Class-based n-gram models of natural language. *Comput. Linguist.* **1992**, *18*, 467–479.
34. Landauer, T.; Foltz, P.; Laham, D. An introduction to latent semantic analysis. *Discourse Process.* **1998**, *25*, 259–284.
35. Lund, K.; Burgess, C.; Atchley, R. Semantic and associative priming in high dimensional semantic space. In Proceedings of the 17th Annual Conference of the Cognitive Science Society, Pittsburgh, PA, USA, 22–25 July 1995; pp. 660–665.
36. Jonnalagadda, S.; Cohen, T.; Wub, S.; Gonzalez, G. Enhancing clinical concept extraction with distributional semantics. *J. Biomed. Inform.* **2012**, *45*, 129–140.
37. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
38. Morin, F.; Bengio, Y. Hierarchical probabilistic neural network language model. In Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, Bridgetown, Barbados, 6–8 January 2005; pp. 246–252.
39. Mnih, A.; Hinton, G. A scalable hierarchical distributed language model. In Proceedings of the 22nd Annual Conference on Neural Information Processing Systems, Vancouver, Canada, 8–11 December 2008; pp. 1081–1088.

40. Huang, E.; Socher, R.; Manning, C.; Ng, A. Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Korea, 8–14 July 2012; pp. 873–882.
41. Mikolov, T.; Karafiát, M.; Burget, L.; Cernocky, J.; Khudanpur, S. Recurrent neural network based language model. In Proceedings of the 11th Annual Conference of the International Speech Communication Association, Makuhari, Japan, 26–30 September 2010; pp. 1045–1048.
42. Natural Language Toolkit. Available online: http://www.nltk.org/ (accessed on 11 December 2015).
43. Ratinov, L.; Roth, D. Design challenges and misconceptions in named entity recognition. In Proceedings of the 13th Conference on Computational Natural Language Learning, Boulder, CO, USA, 4 June 2009; pp. 147–155.
44. Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
45. McCallum, A.; Li, W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the 7th Conference on Natural Language Learning, Geneva, Switzerland, 23–27 August 2003; pp. 188–191.
46. Sutton, C.; McCallum, A.; Rohanimanesh, K. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *J. Mach. Learn. Res.* **2007**, *8*, 693–723.
47. Peng, F.; Feng, F.; McCallum, A. Chinese segmentation and new word detection using conditional random fields. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 23–27 August 2004; pp. 562–568.
48. CRFsuite. Available online: http://www.chokkan.org/software/crfsuite/ (accessed on 11 December 2015).
49. word2vec. Available online: https://code.google.com/p/word2vec/ (accessed on 11 December 2015).
50. GENIA Tagger. Available online: http://www.nactem.ac.uk/tsujii/GENIA/tagger/ (accessed on 11 December 2015).
51. Marcus, M.; Santorini, B.; Marcinkiewicz, M. Building a large annotated corpus of English: The penn treebank. *Comput. Linguist.* **1993**, *19*, 313–330.
52. Sang, E.; Buchholz, S. Introduction to the CoNLL-2000 shared task: Chunking. In Proceedings of the CoNLL-2000, Lisbon, Portugal, 13–14 September 2000; pp. 127–132.
53. DrugBank. Available online: http://www.drugbank.ca/downloads (accessed on 11 December 2015).
54. Drugs@FDA Data Files. Available online: http://www.fda.gov/Drugs/InformationOnDrugs/ucm079750.htm (accessed on 11 December 2015).
55. Leasing Journal Citations (MEDLINE®/PubMed® including OLDMEDLINE). Available online: http://www.nlm.nih.gov/databases/journal.html (accessed on 11 December 2015).
56. Lai, S.; Liu, K.; Xu, L.; Zhao, J. How to generate a good word embedding? **2015**, arXiv:1507.05523.
57. Rocktäschel, T.; Huber, T.; Weidlich, M.; Leser, U. WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 356–363.

58. Grego, T.; Pinto, F.; Couto, F. LASIGE: Using conditional random fields and ChEBI ontology. In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 660–666.
59. Sanchez-Cisneros, D.; Gali, F. UEM-UC3M: An ontology-based named entity recognition system for biomedical texts. In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 622–627.
60. Collazo, A.; Ceballo, A.; Puig, D.; Gutiérrez, Y.; Abreu, J.; Pérez, R.; Orquín, A.; Montoyo, A.; Muñoz, R.; Camara, F. UMCC_DLSI: Semantic and lexical features for detection and classification drugs in biomedical texts. In Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; pp. 636–643.
61. Sang, E.; Meulder, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the 7th Conference on Natural Language Learning, Edmonton, Canada, 31 May–1 June 2003; pp. 142–147.