UMCC_DLSI: Semantic and Lexical features for detection and classification Drugs in biomedical texts

Armando Collazo, Alberto Ceballo, Dennys D. Puig, Yoan Gutiérrez, José I. Abreu, Roger Pérez
DI, University of Matanzas
Autopista a Varadero km 3 ½
Matanzas, Cuba
{armando.collazo, dennys.puig, yoan.gutierrez, jose.abreu, roger.perez}@umcc.cu, alberto.ceballo@infonet.umcc.cu

Antonio Fernández Orquín, Andrés Montoyo, Rafael Muñoz
DLSI, University of Alicante
Carretera de San Vicente
S/N Alicante, Spain
antonybr@yahoo.com, {montoyo, rafael}@dlsi.ua.es

Franc Camara
Independent Consultant
USA
info@franccamara.com

Abstract

In this paper we describe UMCC_DLSI-(DDI) system which attempts to detect and classify drug entities in biomedical texts. We discuss the use of semantic class and words relevant domain, extracted with ISR-WN (Integration of Semantic Resources based on WordNet) resource to obtain our goal. Following this approach our system obtained an F-Measure of 27.5% in the DDIExtraction 2013 (SemEval 2013 task 9).

1. Introduction

To understand biological processes, we must clarify how some substances interact with our body and one to each other. One of these important relations is the drug-drug interactions (DDIs). They occur when one drug interacts with another or when it affects the level, or activity of another drug. DDIs can change the way medications act in the body, they can cause powerful, dangerous and unexpected side effects, and also they can make the medications less effective.

As suggested by (Segura-Bedmar *et al.*, 2011), *"...the detection of DDI is an important research area in patient safety since these interactions can become very dangerous and increase health care costs*". More recent studies (Percha and Altman, 2013) reports that "*...Recent estimates indicate that DDIs cause nearly 74000 emergency room visits and 195000 hospitalizations each year in the USA*".

But, on the other hand, there is an expansion in the volume of published biomedical research, and therefore the underlying biomedical knowledge base (Cohen and Hersh, 2005). Unfortunately, as often happens, this information is unstructured or in the best case scenario semi-structured.

As we can see in (Tari *et al.*, 2010), "*Clinical support tools often provide comprehensive lists of DDIs, but they usually lack the supporting scientific evidences and different tools can return inconsistent results*".

Although, as mentioned (Segura-Bedmar *et al.*, 2011) "*there are different databases supporting healthcare professionals in the detection of DDI, these databases are rarely complete, since their update periods can reach up to three years*". In addition to these and other difficulties, the great amount of drug interactions are frequently reported in journals of clinical pharmacology and technical reports, due to this fact, medical literature becomes most effective source for detection of DDI. Thereby, the management of DDI is a critical issue due to the overwhelming amount of information available on them (Segura-Bedmar *et al.*, 2011).

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 636–643, Atlanta, Georgia, June 14-15, 2013.

1.1. Task Description

With the aim of reducing the time the health care professionals invest on reviewing the literature, we present a feature-based system for drug detection and classification in biomedical texts.

The DDIExtraction2013 task was divided into two subtasks: Recognition and classification of drug names (Task 9.1) and Extraction of drug-drug interactions (Task 9.2). Our system was developed to be presented in the Task 9.1. In this case, participants were to detect and classify the drugs that were present in the test data set which was a set of sentences related to the biomedical domain obtained from a segmented corpus. The output consisted of a list mentioning all the detected drugs with information concerning the sentence it was detected from as well as its offset in that sentence (the position of the first and the last character of the drug in the sentence, 0 being the first character of a sentence). Also the type of the drug should have been provided.

As to the type, participants had to classify entities in one of these four groups[1]:

- Drug: any chemical agent used for treatment, cure, prevention or diagnose of diseases, which have been approved for human usage.
- Brand: any drug which firstly have been developed by a pharmaceutical company.
- Group: any term in the text designating a relation among pharmaceutical substances.
- No-Human: any chemical agent which affects the human organism. An active substance non-approved for human usage as medication.

In the next section of the paper, we present related works (Section 2). In Section 3, we discuss the feature-based system we propose. Evaluation results are discussed in Section 4. Finally, we conclude and propose future work (Section 5).

2. Related Work

One of the most important workshops on the domain of Bioinformatics has been BioCreAtIve (Critical Assessment of Information Extraction in Biology) (Hirschman *et al.*, 2005). This workshop has improved greatly the Information Extraction techniques applied to the biological domain. The goal of the first BioCreAtIvE challenge was to provide a set of common evaluation tasks to assess the state-of-the-art for text mining applied to biological problems. The workshop was held in Granada, Spain on March 28-31, 2004.

According to Hirschman, the first BioCreAtIvE assessment achieved a high level of international participation (27 groups from 10 countries). The best system results for a basic task (gene name finding and normalization), where a balanced 80% precision/recall or better, which potentially makes them suitable for real applications in biology. The results for the advanced task (functional annotation from free text) were significantly lower, demonstrating the current limitations of text-mining approaches.

The greatest contribution of BioCreAtIve was the creation and release of training and test data sets for both tasks (Hirschman *et al.*, 2005).

One of the seminal works where the issue of drug detection was mentioned was (Grönroos *et al.*, 1995). Authors argue the problem can be solved by using a computerized information system, which includes medication data of individual patients as well as information about non-therapeutic drug-effects. Also, they suggest a computerized information system to build decision support modules that, automatically give alarms or alerts of important drug effects other than therapeutic effects. If these warnings concern laboratory tests, they would be checked by a laboratory physician and only those with clinical significance would be sent to clinicians.

Here, it is important to note the appearance of the knowledgebase DrugBank[2]. Since its first release in 2006 (Wishart *et al.*, 2008) it has been widely used to facilitate in silico drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction and general pharmaceutical education. DrugBank has also significantly improved the power and simplicity of its structure query and text query searches.

[1] http://www.cs.york.ac.uk/semeval-2013/task9

[2] http://redpoll.pharmacy.ualberta.ca/drugbank/

Later on, in 2010 Tari propose an approach that integrates text mining and automated reasoning to derive DDIs (Tari *et al.*, 2010). Through the extraction of various facts of drug metabolism, they extract, not only the explicitly DDIs mentioned in text, but also the potential interactions that can be inferred by reasoning. This approach was able to find several potential DDIs that are not present in DrugBank. This analysis revealed that 81.3% of these interactions are determined to be correct.

On the DDIExtraction 2011 (Segura-Bedmar *et al.*, 2011) workshop (First Challenge Task on Drug-Drug Interaction Extraction) the best performance was achieved by the team WBI from Humboldt-Universitat, Berlin. This team combined several kernels and a case-based reasoning (CBR) system, using a voting approach.

In this workshop relation extraction was frequently and successfully addressed by machine learning methods. Some of the more common used features were co-occurrences, character n-grams, Maximal Frequent Sequences, bag-of-words, keywords, etc.

Another used technique is distant supervision. The first system evaluating distant supervision for drug-drug interaction was presented in (Bobić *et al.*, 2012), they have proposed a constraint to increase the quality of data used for training based on the assumption that no self-interaction of real-world objects are described in sentences. In addition, they merge information from IntAct and the University of Kansas Proteomics Service (KUPS) database in order to detect frequent exceptions from the distant supervision assumption and make use of more data sources.

Another important work related to Biomedical Natural Language Processing was BioNLP (Björne *et al.*, 2011) it is an application of natural language processing methods to analyze textual data on biology and medicine, often research articles. They argue that information extraction techniques can be used to mine large text datasets for relevant information, such as relations between specific types of entities.

Inspired in the previews works the system we propose makes use of machine learning methods too, using some of the common features described above, such as the n-grams and keywords and co-occurrences, but we also add some semantic information to enrich those features.

3. System Description

As it has been mentioned before, the system was developed to detect and classify drugs in biomedical texts, so the process is performed in two main phases:

- drug detection.
- drug classification.

Both phases are determined by the following stages, described in Figure 1:

I. Preprocessing
II. Feature extraction
III. Classification

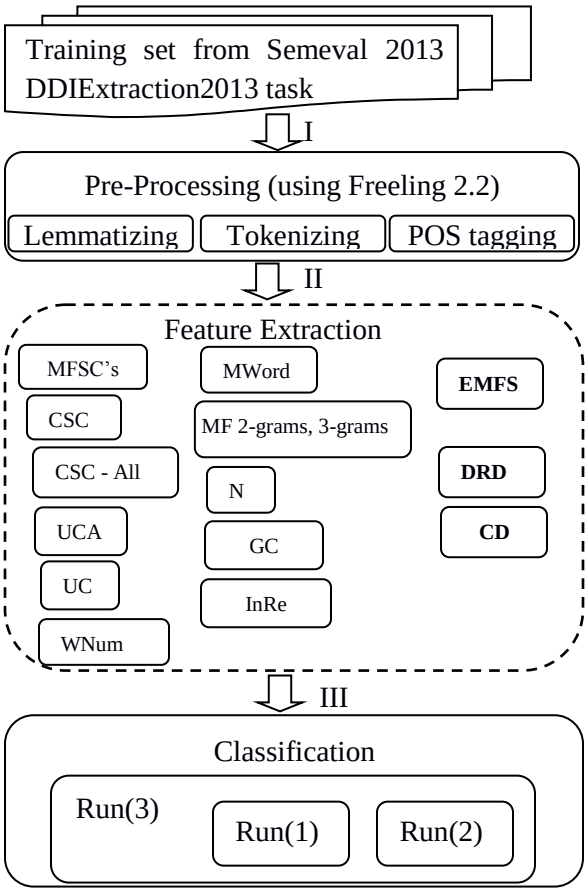

Figure 1. Walkthrough system process.

Given a biomedical sentence, the system obtains the lemmas and POS-tag of every token

of the sentence, by means of Freeling tool[3]. After that, it is able to generate candidates according to certain parameters (see section 3.3).

Then, all the generated candidates are processed to extract the features needed for the learning methods, in order to determine which candidates are drugs.

After the drugs are detected, the system generates a tagged corpus, following the provided training corpus structure, containing the detected entities, and then it proceeds to classify each one of them. To do so, another supervised learning algorithm was used (see section 3.3).

3.1. Candidates generation

Drugs and drug groups, as every entity in Natural Language, follow certain grammatical patterns. For instance, a drug is usually a noun or a set of nouns, or even a combination of verbs and nouns, especially verbs in the past participle tense and gerunds. But, one thing we noticed is that both drugs and drug groups end with a noun and as to drug groups that noun is often in the plural.

Based on that idea, we decided to generate candidates starting from the end of each sentence and going forward.

Generation starts with the search of a pivot word, which in this case is a noun. When the pivot is found, it is added to the candidates list, and then the algorithm takes the word before the pivot to see if it complies with one of the patterns i.e. if the word is a noun, an adjective, a gerund or past participle verb. If it does, then it and the pivot form another candidate.

After that, the algorithm continues until it finds a word that does not comply with a pattern. In this case, it goes to the next pivot and stops when all the nouns in the sentence have been processed, or the first word of the sentence is reached.

3.2. Feature Description

For the DDIExtraction2013[4] task 9 three runs of the same system were performed with different features each time. The next sections describes the features we used.

3.2.1. Most Frequent Semantic Classes (MFSC)

Given a word, its semantic class label (Izquierdo *et al.*, 2007) is obtained from WordNet using the ISR-WN resource (Gutiérrez *et al*., 2011; 2010). The semantic class is that associated to the most probable sense of the word. For each entity in the training set we take the words in the same sentence and for each word its semantic class is determined. This way, we identify the 400[5] most frequent semantic classes associated to words surrounding the entities in the training set.

For a candidate entity we use 400 features to encode information with regard to whether or not in its same sentence a word can be found belonging to one of the most frequent semantic classes.

Each one of these features takes a value representing the distance (measured in words) a candidate is from the nearest word with same semantic class which represents the attribute.

If the word is to the left of the candidate, the attribute takes a negative value, if it is to the right, the value is positive, and zero if no word with that semantic class is present in the sentence the candidate belongs to.

To better understand that, consider A1 is the attribute which indicates if in the sentence of the candidate a word can be found belonging to the semantic class 1. Thus, the value of A1 is the distance the candidate is from the closest word with semantic class 1 in the sentence that is being analyzed.

3.2.2. Candidate Semantic Class (CSC)

The semantic class of candidates is also included in the feature set, if the candidate is a multi-word, then the semantic class of the last word (the pivot word) is taken.

[3] http://nlp.lsi.upc.edu/freeling/

[4] http://www.cs.york.ac.uk/semeval-2013/task9/

[5] This value was extracted from our previous experiment.

3.2.3. Most Frequent Semantic Classes from Entities (EMFSC)

In order to add more semantic information, we decided to find the most frequent semantic classes among all the entities that were tagged in the training data set. We included, in the feature set, all the semantic classes with a frequency of eight or more, because all the classes we wanted to identify were represented in that threshold. In total, they make 29 more features. The values of every one of them, is the sum of the number of times it appears in the candidate.

3.2.4. Candidate Semantic Class All Words (CSC-All)

This feature is similar to CSC, but in this case the candidate is a multi-word, we not only look for the semantic class of the pivot, but also the whole candidate as one.

3.2.5. Drug-related domains (DRD)

Another group of eight attributes describes how many times each one of the candidates belongs to one of the following drug-related domains (DRD) (medicine, anatomy, biology, chemistry, physiology, pharmacy, biochemistry, genetics). These domains where extracted from WordNet Domains. In order to determine the domain that a word belongs to, the proposal of DRelevant (Vázquez *et al.*, 2007; Vázquez *et al.*, 2004) was used.

To illustrate how the DRD features take their values, consider the following sentence:

"…until the lipid response to Accutane is established."

One of the candidates the system generates would be "lipid response". It is a two-word candidate, so we take the first word and see if it belongs to one of the above domains. If it does, then we add one to that feature. If the word does not belong to any of the domains, then its value will be zero. We do the same with the other word. In the end, we have a collection where every value corresponds to each one of the domains. For the example in question the collection would be:

medicine	1
anatomy	0
biology	0
chemistry	0
physiology	1
pharmacy	0
biochemistry	0
genetics	0

Table 1. DRD value assignment example.

3.2.6. Candidate word number (WNum)

Because there are candidates that are a multi-word and others that are not, it may be the case that a candidate, which is a multi-word, has an EMFSC bigger than others which are not a multi-word, just because more than one of the words that conform it, have a frequent semantic class.

We decided to add a feature, called WNum, which would help us normalize the values of the EMFSC. The value of the feature would be the number of words the candidate has. Same thing happens with DRD.

3.2.7. Candidate Domain (CD)

The value of this nominal feature is the domain associated to the candidate. If the candidate is a multi-word; we get the domain of all the words as a whole. In both cases the domain for a single word as well as for a multi-word is determined using the relevant domains obtained by (Vázquez et al., 2007; Vázquez et al., 2004).

3.2.8. Maximum Frequent 2-grams, 3-grams

Drugs usually contain sequences of characters that are very frequent in biomedical domain texts. These character sequences are called *n*-grams, where *n* is the number of characters in the sequence. Because of that, we decided to add the ten most frequent *n*-grams with *n* between two and three. The selected *n*-grams are the following: "in" (frequency: 8170), "ne" (4789), "ine" (3485), "ti" (3234), "id" (2768), "an" (2704), "ro" (2688), "nt" (2593), "et" (2423), "en" (2414).

These features take a value of one if the candidate has the corresponding character sequence and zero if it does not. For instance: if

we had the candidate "panobinostat" it will generate the following collection:

"in"	1
"ne"	0
"ine"	0
"ti"	0
"id"	0
"an"	1
"ro"	0
"nt"	0
"et"	0
"en"	0

Table 2. MF 2-gram, 3-gram.

3.2.9. Uppercase (UC), Uppercase All (UCA). Multi-word (MWord) and Number (N)

Other features say if the first letter of the candidate is an uppercase; if all of the letters are uppercase (UCA); if it is a multi-word (MWord) and also if it is in the singular or in the plural (N).

3.2.10. L1, L2, L3 and R1, R2, R3

The Part-of-Speech tags of the closest three surrounding words of the candidates are also included. We named those features L1, L2, and L3 for POS tags to the left of the candidate, and R1, R2, and R3 for those to the right.

3.2.11. POS-tagging combination (GC)

Different values are assigned to candidates, in order to identify its POS-tagging combination. For instance: to the following entity "combined oral contraceptives" taken from DDI13-train-TEES-analyses-130304.xml[6] training file, which was provided for task 9.1, corresponds 5120. This number is the result of combining the four grammatical categories that really matter to us: R for adverb, V for verb, J for adjective, N for noun.

A unique number was given to each combination of those four letters. We named this feature GC.

[6] http://www.cs.york.ac.uk/semeval-2013/task9

3.2.12. In resource feature (InRe)

A resource was created which contains all the drug entities that were annotated in the training corpus, so another attribute tells the system if the candidate is in the resource.

Since all of the entities in the training data set were in the resource this attribute could take a value of one for all instances. Thus the classifier could classify correctly all instances in the training data set just looking to this attribute, which is not desirable. To avoid that problem, we randomly set its value to zero every 9/10 of the training instances.

3.3. Classification

All the features extracted in the previous stages are used in this stage to obtain the two models, one for drug detection phase, and the other for drug classification phase.

We accomplished an extensive set of experiments in order to select the best classifier. All algorithms implemented in WEKA, except those that were designed specifically for a regression task, were tried. In each case we perform a 10-fold cross-validation. In all experiments the classifiers were settled with the default configuration. From those tests we select a decision tree, the C4.5 algorithm (Gutiérrez *et al.*, 2011; 2010) implemented as the J48 classifier in WEKA. This classifier yields the better results for both drug detection and drug classification.

The classifier was trained using a set of 463 features, extracted from the corpus provided by SemEval 2013, the task 9 in question.

As it was mentioned before, three runs were performed for the competition. Run (1) used the following features for drug detection: MFSC (only 200 frequent semantic classes), MF 2-grams, 3-grams, UC, UCA, MWord, N, L1, L2, L3, R1, R2, R3, CSC, CD, WNum, GC and InRe.

Drug classification in this run used the same features except for CD, WNum, and GC. Run (2) has all the above features, but we added the remaining 200 sematic classes that we left out in Run (1) to the detection and the classification models. In Run (3), we added EMFSC feature to the detection and the classification models.

4. Results

In the task, the results of the participants were compared to a gold-standard and evaluated according to various evaluation criteria:

- Exact evaluation, which demands not only boundary match, but also the type of the detected drug has to be the same as that of the gold-standard.
- Exact boundary matching (regardless of the type).
- Partial boundary matching (regardless of the type)
- Type matching.

Precision and recall were calculated using the scoring categories proposed by MUC [7]:

- COR: the output of the system and the gold-standard annotation agree.
- INC: the output of the system and the gold-standard annotation disagree.
- PAR: the output of the system and the gold-standard annotation are not identical but has some overlapping text.
- MIS: the number of gold-standard entities that were not identify by the system.
- SPU: the number of entities labeled by the system that are not in the gold-standard.

Table 3 , Table 4 and Table 5 show the system results in the DDIExtraction2013 competition for Run (1).

Run (2) and Run (3) results are almost the same as Run (1). It is an interesting result since in those runs 200 additional features were supplied to the classifier. In feature evaluation, using CfsSubsetEval and GeneticSearch with WEKA we found that all these new features were ranked as worthless for the classification. On the other hand, the following features were the ones that really influenced the classifiers: MFSC (215 features only), MF 2-grams, 3-grams ("ne", "ine", "ti", "ro", "et", "en"), WNum, UC, UCA, L1, R1, CSC, CSC-All, CD, DRD (anatomy, physiology, pharmacy, biochemistry), InRe, GC and EMFS, specifically music.n.01, substance.n.01, herb.n.01, artifact.n.01, nutriment.n.01, nonsteroidal_anti-inflammatory.n.01, causal_agent.n.01 have a frequency of 8, 19, 35, 575, 52, 80, 63 respectively.

Measure	Strict	Exact Matching	Partial Matching	Type
COR	319	354	354	388
INC	180	145	0	111
PAR	0	0	145	0
MIS	187	187	187	187
SPU	1137	1137	1137	1137
Precision	0.19	0.22	0.22	0.24
Recall	0.47	0.52	0.62	0.57

Table 3. Run (1), all scores.

Measure	Drug	Brand	Group	Drug_n
COR	197	20	93	9
INC	23	2	43	1
PAR	0	0	0	0
MIS	131	37	19	111
SPU	754	47	433	14
Precision	0.2	0.29	016	0.38
Recall	0.56	0.34	0.6	0.07
F1	0.3	0.31	0.26	0.12

Table 4. Scores for entity types, exact matching in Run (1).

	Precision	**Recall**	**F1**
Macro average	0.26	0.39	0.31
Strict matching	0.19	0.46	0.27

Table 5. Macro average and Strict matching measures in Run (1).

5. Conclusion and future works

In this paper we show the description of UMCC_DLSI-(DDI) system, which is able to detect and classify drugs in biomedical texts with acceptable efficacy. It introduces in this thematic the use of semantic information such as semantic classes and the relevant domain of the words, extracted with ISR-WN resource. With this approach we obtained an F-Measure of 27.5% in the Semeval DDI Extraction2013 task 9.

As further work we propose to eliminate some detected bugs (i.e. repeated instances, multiwords missed) and enrich our knowledge base (ISR-WN), using biomedical sources as UMLS[8], SNOMED[9] and OntoFis[10].

[7] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_sw/muc_sw_manual.html

[8] http://www.nlm.nih.gov/research/umls

[9] http://www.ihtsdo.org/snomed-ct/

[10] http://rua.ua.es/dspace/handle/10045/14216

Acknowledgments

This research work has been partially funded by the Spanish Government through the project TEXT-MESS 2.0 (TIN2009-13391-C04), "Análisis de Tendencias Mediante Técnicas de Opinión Semántica" (TIN2012-38536-C03-03) and "Técnicas de Deconstrucción en la Tecnologías del Lenguaje Humano" (TIN2012-31224); and by the Valencian Government through the project PROMETEO (PROMETEO/2009/199).

References

Björne, J.; A. Airola; T. Pahikkala and T. Salakoski Drug-Drug Interaction Extraction from Biomedical Texts with SVM and RLS Classifiers Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction, 2011, 761: 35-42.

Bobić, T.; R. Klinger; P. Thomas and M. Hofmann-Apitius Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactions EACL 2012, 2012: 35.

Cohen, A. M. and W. R. Hersh A survey of current work in biomedical text mining Briefings in bioinformatics, 2005, 6(1): 57-71.

Grönroos, P.; K. Irjala; J. Heiskanen; K. Torniainen and J. Forsström Using computerized individual medication data to detect drug effects on clinical laboratory tests Scandinavian Journal of Clinical & Laboratory Investigation, 1995, 55(S222): 31-36.

Gutiérrez, Y.; A. Fernández; A. Montoyo and S. Vázquez. Integration of semantic resources based on WordNet. XXVI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, Universidad Politécnica de Valencia, Valencia, SEPLN 2010, 2010. 161-168 p. 1135-5948

Gutiérrez, Y.; A. Fernández; A. Montoyo and S. Vázquez Enriching the Integration of Semantic Resources based on WordNet Procesamiento del Lenguaje Natural, 2011, 47: 249-257.

Hirschman, L.; A. Yeh; C. Blaschke and A. Valencia Overview of BioCreAtIvE: critical assessment of information extraction for biology BMC bioinformatics, 2005, 6(Suppl 1): S1.

Izquierdo, R.; A. Suárez and G. Rigau A Proposal of Automatic Selection of Coarse-grained Semantic Classes for WSD Procesamiento del Lenguaje Natural, 2007, 39: 189-196.

Percha, B. and R. B. Altman Informatics confronts drug–drug interactions Trends in pharmacological sciences, 2013.

Segura-Bedmar, I.; P. Martínez and D. Sánchez-Cisneros The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts Challenge Task on Drug-Drug Interaction Extraction, 2011, 2011: 1-9.

Tari, L.; S. Anwar; S. Liang; J. Cai and C. Baral Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism Bioinformatics, 2010, 26(18): i547-i553.

Vázquez, S.; A. Montoyo and Z. Kozareva. Extending Relevant Domains for Word Sense Disambiguation. IC-AI'07. Proceedings of the International Conference on Artificial Intelligence USA, 2007.

Vázquez, S.; A. Montoyo and G. Rigau. Using Relevant Domains Resource for Word Sense Disambiguation. IC-AI'04. Proceedings of the International Conference on Artificial Intelligence, Ed: CSREA Press. Las Vegas, E.E.U.U., 2004.

Wishart, D. S.; C. Knox; A. C. Guo; D. Cheng; S. Shrivastava; D. Tzur; B. Gautam and M. Hassanali DrugBank: a knowledgebase for drugs, drug actions and drug targets Nucleic acids research, 2008, 36(suppl 1): D901-D906.