

Drug-Drug Interaction Detection: A New Approach Based on Maximal Frequent Sequences

Sandra García-Blasco Roxana Danger Paolo Rosso

Natural Language Engineering Lab. - ELiRF
DSIC - Univ. Politécnic Valencia
sangarbl@posgrado.upv.es
{rdanger,prossso}@dsic.upv.es

BioSEPLN, September 2010

Outline

- 1 Introduction
 - The problem: Drug-Drug Interaction Detection
 - Approximations
- 2 Out Proposal
 - Method Proposed
 - The Algorithm
- 3 Experimentation
 - Corpus and preprocessing
 - Results
- 4 Conclusions

Outline

- 1 Introduction
 - The problem: Drug-Drug Interaction Detection
 - Approximations
- 2 Out Proposal
 - Method Proposed
 - The Algorithm
- 3 Experimentation
 - Corpus and preprocessing
 - Results
- 4 Conclusions

What are Drug-Drug Interactions?

- A **drug-drug interaction (DDI)** occurs when the effects of a drug are modified by the presence of other drugs.
- Its consequences may be very harmful for the patient's health and could even cause his dead.
- This gives us an idea of how important is for health-care professionals to keep their databases up-to-date with new DDI.

What are Drug-Drug Interactions?

- A **drug-drug interaction (DDI)** occurs when the effects of a drug are modified by the presence of other drugs.
- Its consequences may be very harmful for the patient's health and could even cause his dead.
- This gives us an idea of how important is for health-care professionals to keep their databases up-to-date with new DDI.

What are Drug-Drug Interactions?

- Most of the new discoveries in DDI are published in bibliographic databases on health and biomedicine, like MEDLINE:
 - MEDLINE has over 18 million references to journal articles
 - In 2009, over 712.000 articles added.
- This growing amount of information leaves very clear how necessary is to find efficient methods that help health-care professionals to better deal with all this information.

What are Drug-Drug Interactions?

- Most of the new discoveries in DDI are published in bibliographic databases on health and biomedicine, like MEDLINE:
 - MEDLINE has over 18 million references to journal articles
 - In 2009, over 712.000 articles added.
- This growing amount of information leaves very clear how necessary is to find efficient methods that help health-care professionals to better deal with all this information.

Outline

- 1 Introduction
 - The problem: Drug-Drug Interaction Detection
 - **Approximations**
- 2 Out Proposal
 - Method Proposed
 - The Algorithm
- 3 Experimentation
 - Corpus and preprocessing
 - Results
- 4 Conclusions

Approximations of other authors

In (Segura-Bedmar, 2010) two different techniques for DDI detection are presented:

- A hybrid approach, combining **shallow parsing** and **pattern matching**. The patterns used in this technique were described by a pharmacist, and they obtained 48.7% precision, and 25.7% recall.
- An approach based on a supervised machine learning approach, specifically **kernel methods**, obtaining 55% precision and 84% recall.

Our proposal

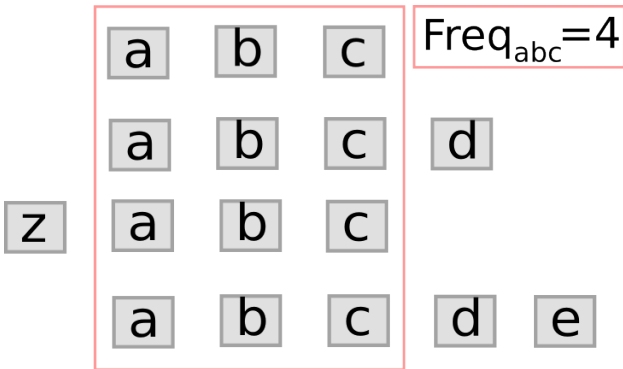
- Objective: Automatically determining the patterns that identify DDI from a set of documents.
- Our hypothesis holds that there must be patterns that we will find repeated if we look through a large amount of biomedical texts, and those patterns will help to identify new drug drug interactions.
- The method proposed in this paper is **language** and **domain independent**.

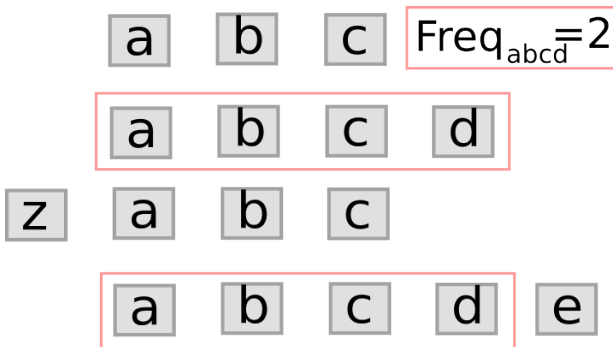
Outline

- 1 Introduction
 - The problem: Drug-Drug Interaction Detection
 - Approximations
- 2 **Out Proposal**
 - **Method Proposed**
 - The Algorithm
- 3 Experimentation
 - Corpus and preprocessing
 - Results
- 4 Conclusions

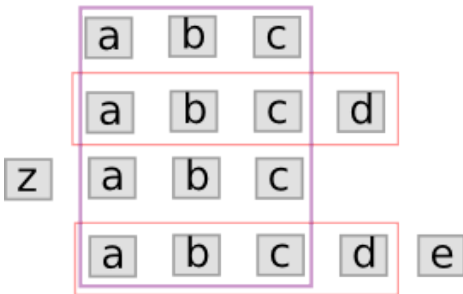
Maximal Frequent Sequences

- A **sequence** is an ordered list of elements, i.e. words.
- The **frequency** of a sequence is the number of times that the sequence appears.
- A sequence will be **β -frequent** if it is included in β sentences
- A sequence R is **subsequence** of a sequence T if all the elements of R appear in T in the same order. For example:
 - If $R = \langle abcde \rangle$ and $T = \langle bcd \rangle$ then,
 T is subsequence of R
- A **maximal sequence** is a sequence that is not a subsequence of any other.





Sequences with Freq > 1

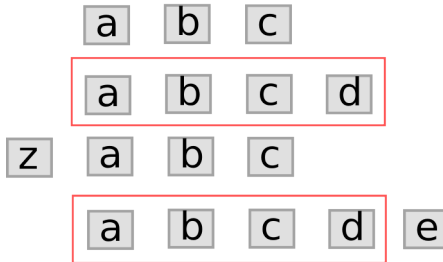


Maximal Frequent Sequences

Definition

Maximal Frequent Sequences (MFS) will be all the sequences that are frequent and that are not subsequence of any other.

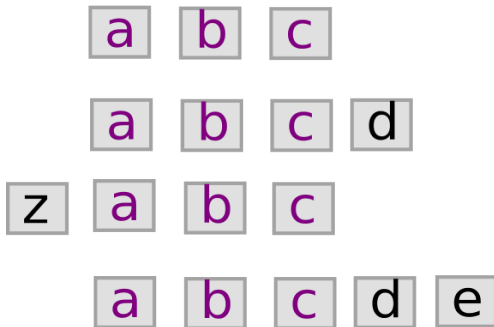
Maximal Sequences with Freq > 1



Gap

- In order to make this maximal frequent sequences more flexible, the concept of **gap** is introduced (Garcia-Hernandez, 2007)
- The **gap** is the maximum distance that is allowed between two words of a MFS. With a $gap = 0$, the words in the MFS will be adjacent words in the original text.

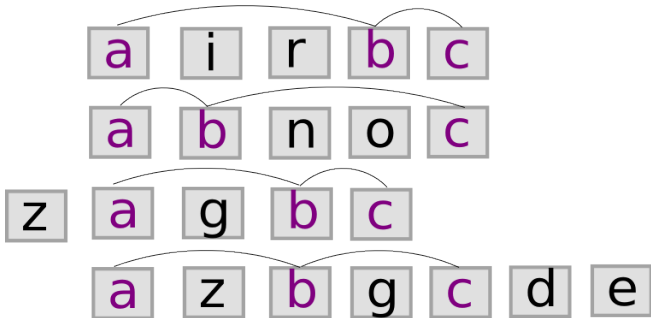
GAP



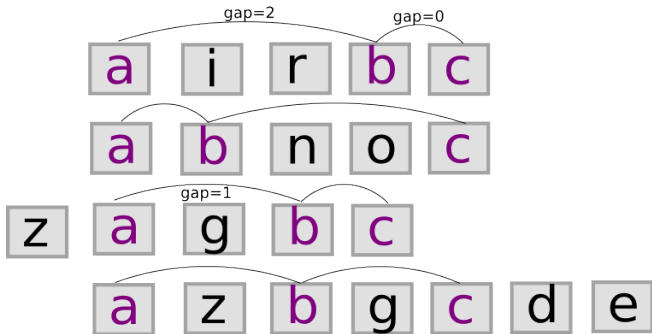
GAP



GAP



GAP



Outline

- 1 Introduction
 - The problem: Drug-Drug Interaction Detection
 - Approximations
- 2 Out Proposal
 - Method Proposed
 - **The Algorithm**
- 3 Experimentation
 - Corpus and preprocessing
 - Results
- 4 Conclusions

The Algorithm

- The algorithm presented is based on the *Apriori Algorithm* (Agrawal and Srikant, 1994), but with the difference that our algorithm takes into account the sequentiality of the elements, i.e. words, allowing gaps between them.
- The algorithm can be divided into 3 stages:
 - 1 Getting bag of words
 - 2 Finding candidates
 - 3 Merging Patterns

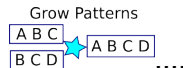
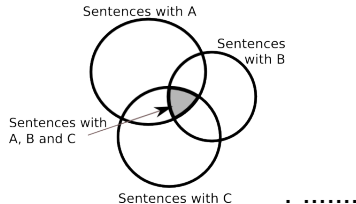
The algorithm

- The algorithm has three parameters:
 - `minFreq` minimum number of sentences where the *MFS* should appear.
 - `minLength` minimum length of the *MFS*.
 - `gap` maximum distance allowed between two words of the *MFS*.

The Algorithm - An overview

Input: minFreq, minLength, gap

- 1 Build a Bag of Words with the frequent words
- 2 Combinations of length 3 of frequent words
- 3 For each combination:
 - If $\text{size}(\text{intersection}) < \text{minFreq}$, discard
- 4 Permute combinations
- 5 For each permutation:
 - If $\#\text{sent with perm in right order} < \text{minFreq}$, discard
- 6 Grow Patterns to make them Maximal
- 7 Remove patterns with length $< \text{minLength}$



Outline

- 1 Introduction
 - The problem: Drug-Drug Interaction Detection
 - Approximations
- 2 Out Proposal
 - Method Proposed
 - The Algorithm
- 3 **Experimentation**
 - **Corpus and preprocessing**
 - Results
- 4 Conclusions

Corpus

- The **DrugDDI** corpus (Segura-Bedmar, 2010) is a drug-drug interaction corpus annotated with linguistic information, named entities and drug interactions.
- Drugs are tagged in the corpus, according to their type. There are 6 types:
 - Clinical drug (clnd)
 - Pharmacological Substance (phsu)
 - Antibiotic (antb)
 - Biologically Active Substance (bacs)
 - Chemical viewed structurally (chvs)
 - Amino acid, Peptide or Protein (aapp)

- The corpus consists of 579 documents from the DrugBank database, with an average of 10.3 sentences and 5.46 interactions per document.
- The corpus has been divided into two sets:
 - Training with 446 documents.
 - Test with 133 documents.

Preprocess

Three different versions of the corpus were obtained

Normal Original Text

- Acetazolamida may increase the effects of other folic acid antagonists

6Drug Each drug name was substituted by its type

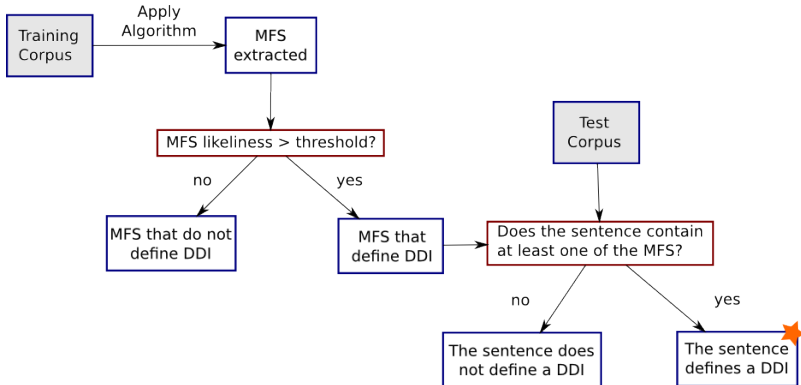
- phsu may increase the effects of other phsu

#Drug# Each drug name was substituted by #drug#

- #drug# may increase the effects of other #drug#

Experiments

Objective Identify drug drug interactions in biomedical texts using *maximal frequent sequences*.



Experiments

- First, the algorithm is used to extract **MFS** from the training set using the following configurations:

`minLength` 4
`minFreq` 10, 15, 20
`gap` 0, 1, 2

- Each experiment was repeated for each one of the 3 versions of the corpus: *norm*, *6drugs*, *#drug#*.

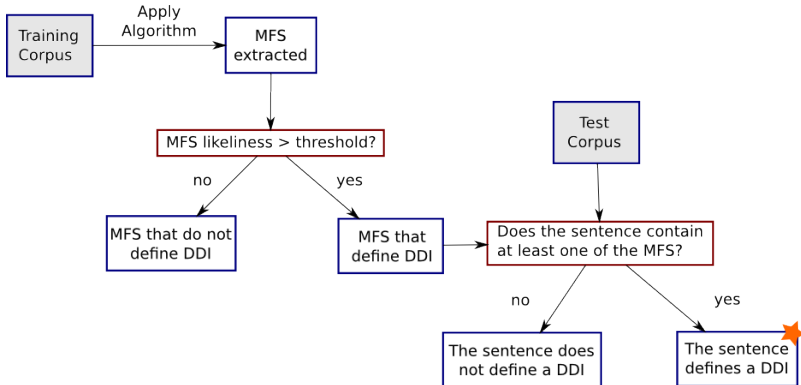
Experiments

- Next, the **MFS** detected where rated using a new function that we define, **likeliness**, that is the probability of the MFS to describe a DDI. Likeliness is defined as:

$$\textit{likeliness}(MFS_i) = \frac{\text{times } MFS_i \text{ identifies DDI}}{\text{times } MFS_i \text{ appears}}$$

Experiments

Objective Identify drug drug interactions in biomedical texts using *maximal frequent sequences*.

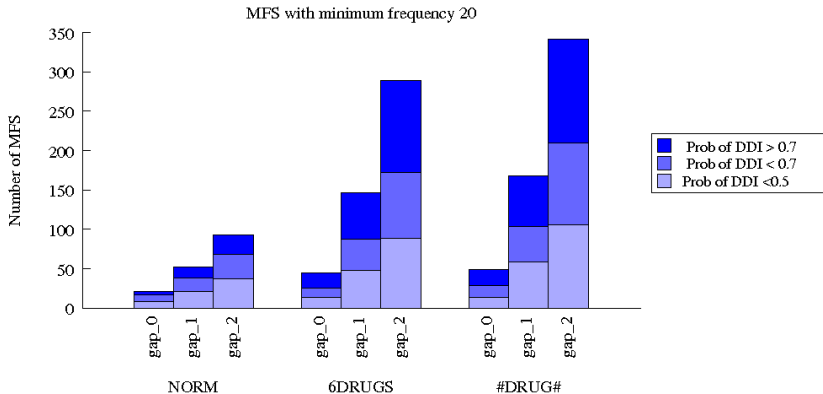


Outline

- 1 Introduction
 - The problem: Drug-Drug Interaction Detection
 - Approximations
- 2 Out Proposal
 - Method Proposed
 - The Algorithm
- 3 **Experimentation**
 - Corpus and preprocessing
 - **Results**
- 4 Conclusions

Results

- The algorithm has detected maximal frequent sequences that describe drug-drug interaction.



Results

Example (MFS)

(' #drug#', 'may', 'the', 'effects', 'of', ' #drug#')

Extracted from sentences like:

- *Acetazolamide **may** increase **the effects of** other folic acid antagonists*
- *Alcohol **may** potentiate **the side effects of** bromocriptine mesylate*
- *Dopamine D2 receptor antagonists (e.g., phenothiazines, butyrophenones, risperidone) and isoniazid **may** reduce **the therapeutic effects of** levodopa*
- *Concomitant administration of other sympathomimetic agents **may** potentiate **the undesirable effects of** FORADIL*

Using #drug#, minFreq = 10 and gap = 1

Examples of the MFS extracted

MFS description	Sample	<i>freq</i>	<i>likeliness</i>
With verbs denoting effects	('drug#', 'may', 'increase', 'of')	30	0.93
	('may', 'decrease', 'the', 'of')	21	0.90
	('drug#', 'may', 'enhance', 'the', 'of')	10	1.0
	('drug#', 'is', 'administered', 'with')	21	0.81
With 2 or more drugs	('drug#', 'may', 'the', 'effects', 'drug#')	13	1.0
	('drug#', 'should', 'not', 'be', 'with', 'drug#')	11	1.0
	('drug#', 'reduce', 'the', 'of', 'drug#')	15	0.93

Results

To calculate the performance of the method the measures of precision, recall and F_1 -measure are used.

Precision is defined as the number of sentences describing DDI retrieved divided by the total number of sentences retrieved.

Recall is defined as the number of sentences describing DDI retrieved divided by the total number of existing sentences describing DDI.

F_1 -measure is the harmonic mean of precision and recall.

Results

- The baseline is the one given by tagging all the sentences as DDI.

	Precision	Recall	F_1
<i>baseline</i>	0.40	1	0.28
<i>norm</i>	0.68	0.41	0.51
<i>6drugs</i>	0.48	0.93	0.63
<i>#drug#</i>	0.46	0.95	0.62

Table: Comparison of Results




- As the table shows, some of the parameters give a very high recall value (95%).

Conclusions I



- DDIs are described by the researchers using a reduced vocabulary and similar sentences structures are used to describe drug-drug interactions.
- Maximal Frequent Sequences are able to extract repeated patterns and has been proved to be a good method for drug-drug interaction detection.
- The method proposed is **domain** and **language independent** , and can be applied in many other tasks, like Protein-Protein or Protein-Drug Interaction detection.
- This method does not require any domain specific knowledge, extracting the patterns directly from a sample corpus.

Thank you.

References I

-  Rakesh Agrawal and Ramakrishnan Srikant.
Fast algorithms for mining association rules in large databases.
In *VLDB*, pages 487–499, 1994.
-  Helena Ahonen-Myka.
Finding all maximal frequent sequences in text, 1999.
-  Helena Ahonen-Myka.
Discovery of frequent word sequences in text.
In *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery*, pages 180–189, London, UK, 2002. Springer-Verlag.

References II

-  Rosa M. Coyotl-Morales, Luis Villaseñor-Pineda, Manuel Montes-y Gómez, and Paolo Rosso.
Authorship attribution using word sequences.
In *LNCS*, pages 844–853. Springer, 2006.
-  René A. García.
Algoritmos para el descubrimiento de patrones secuenciales maximales.
PhD thesis, INAOE. Mexico, September 2007.

References III



Sandra García-Blasco.

Extracción de secuencias maximales de una colección de textos.

Final degree project, ETSInf, Universidad Politécnica de Valencia, Spain, December 2009.



Isabel Segura-Bedmar.

Application of Information Extraction techniques to pharmacological domain: Extracting drug-drug interactions.

PhD thesis, Universidad Carlos III, Madrid, Spain, April 2010.