

Spanish resources of TrendMiner Project

Course on Advanced Topics

Combining Language and Web Technologies,

UNED, January 22, 2015

Paloma Martínez

Advanced Databases Group

labda.inf.uc3m.es

Universidad Carlos III de Madrid



CONTENTS

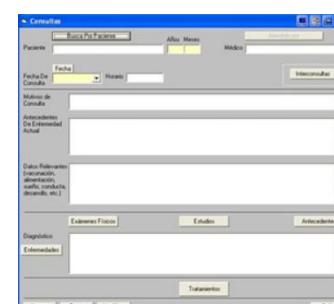
1. Challenges in automatic semantic analysis of health information
2. Objective
3. Resources
4. Linguistic Processor
5. Real-time prototype
6. Annotation pipeline evaluation
7. Other methods to extract drug-effect relations
8. Possible extensions

RETOS EN EL ANÁLISIS SEMÁNTICO AUTOMÁTICO DE INFORMACIÓN DE SALUD

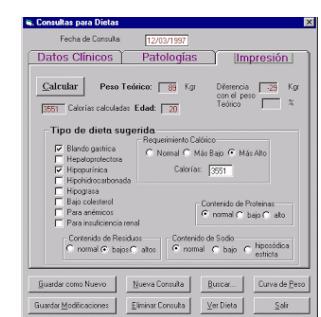
- Extracción y Recuperación de información en el dominio biomédico en distintos medios (publicaciones científicas, redes sociales, notas clínicas)
- ¿Cuántos datos estructurados se procesan de la Historia Clínica Electrónica? ¿Y con los no estructurados, qué se hace?
- Aplicaciones:
 1. Soporte a la codificación ICD9/10, SNOMED CT,(p.e. diagnósticos en partes de alta en urgencias)
 2. Filtrado de grandes volúmenes de información
 3. Extracción de información para alimentar BD (p.e. extracción de interacciones entre fármacos a partir de literatura médica)
 4. Monitorización de eventos médicos en distintos medios



No Estructurados



Estructurados



RETOS EN EL ANÁLISIS SEMÁNTICO AUTOMÁTICO DE INFORMACIÓN DE SALUD

Enfermedades

SNOMED

CIE-10

Entender el lenguaje
de la Salud

MedDRA

TRATAMIENTOS

SÍNTOMAS

Fármacos

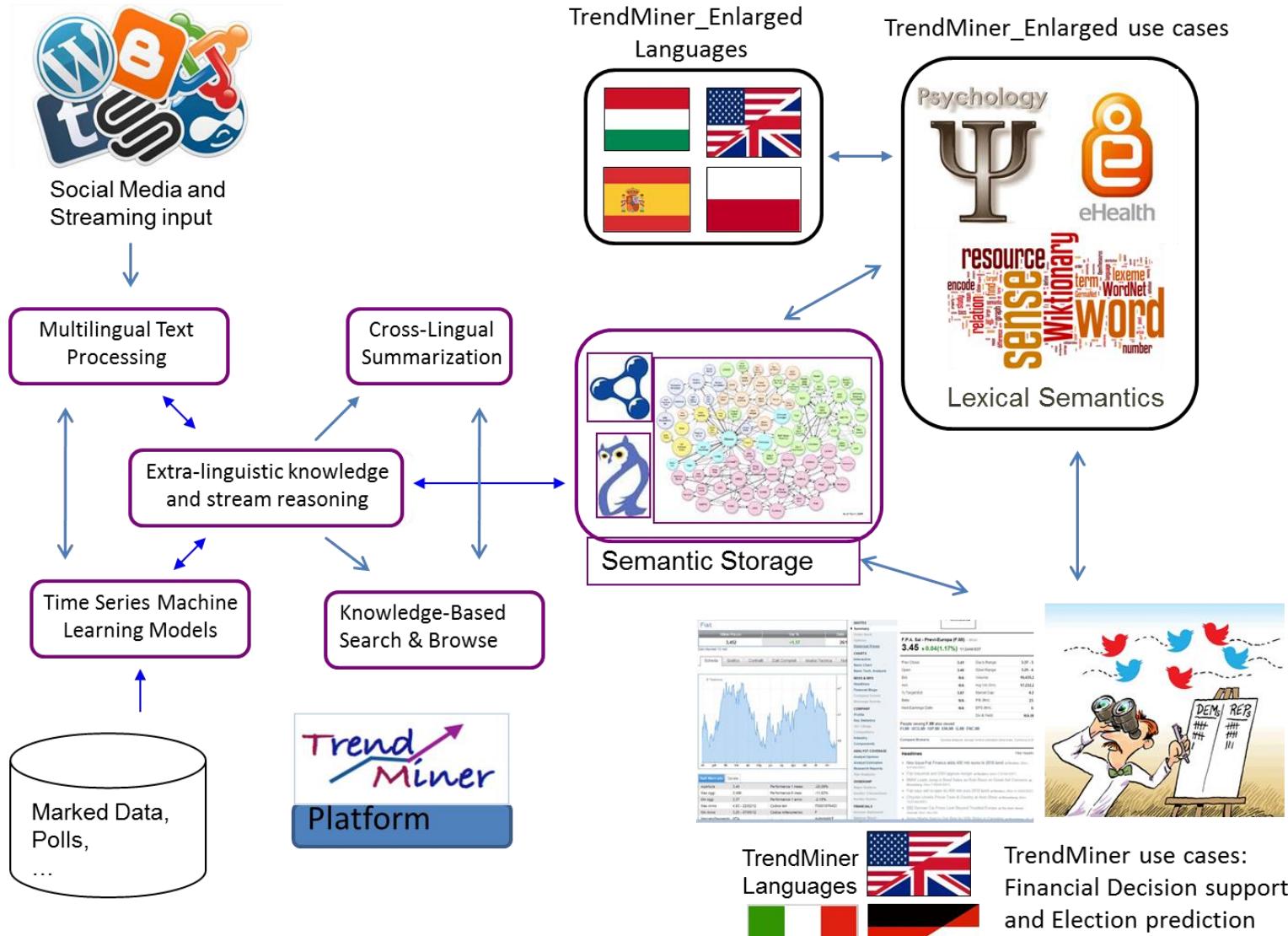
1. Español
2. ≠ tipos de lenguaje (orientado a pacientes, científico, clínico)
3. Fenómenos propios (abreviaturas, gran ambigüedad, ...)

Principios
activos



Trend Miner

Large-scale
Cross-lingual Trend Mining
Summarization



OBJECTIVE

- To detect drugs and medical events mentions (drugs, diseases, symptoms, Adverse effects,) from social media.
- Social medial sources can be valuable sources monitoring medical events
- Different applications:
 - Pharmacovigilance tasks performed in medicines agencies and pharma companies.
 - Filtering, classification and monitoring of health-related social networks (blogs, forums,)
 - Information Extraction tasks

RESOURCES

(1) Analyzed and monitored sources



Spanish patient Forums



RESOURCES

(1) Analyzed and monitored sources

Example of post in Forumclinic

Source Texts

Daedalus_UC3M_Trendminer [2014-09-03T16:13:48 GMT]

hola lola, al leer tu mensaje me he sentido identificada contigo. a mi tambien me prescribieron el taxotere y la herceptin@. te explico: el 2 de marzo fue mi primera sesion, me estrenaba en todo este mundo... la herceptina la tolere bien pero el taxotere al poco de entrar en vena me dio un eritema cutaneo y me lo quitaron enseguida.

+ taxotere

+ eritema

+ taxotere -> eritema [adverseEffect]

RESOURCES

(2) Integrated existing semantic domain resources



MedDRA



35.259 terms

nauseas estomago revuelto | sentirse
mareado | nauseas | nauseas
solas | nauseoso | nauseoso | ansia
nauseosa |



16.418 drugs and 2.228
active substances



Ibuprofeno
algiasdin | aprofeno | aragel | articalm | ast
efor | brufen | dalsy | dersindol | diltix | dole
ncar | doltra | espididol | espidifen |

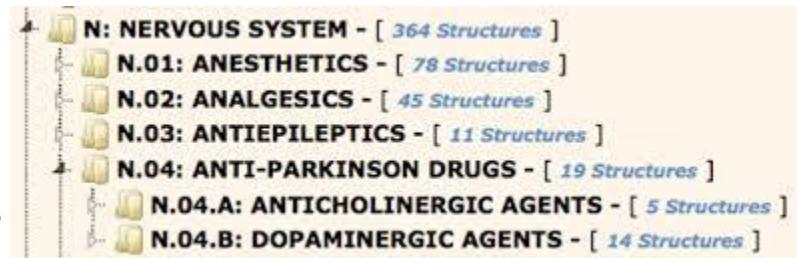
Anatomical Therapeutic Chemical (ATC) Classification System



World Health
Organization



2.566 ATC codes



Unified Medical Language System® (UMLS®)



42.548 main diseases

Cáncer | neoplasia maligna |

RESOURCES

(2) Integrated existing semantic domain resources

ATC Structure

ATC is a system of alphanumeric codes developed by the WHO for the classification of drugs and other medical products

Level 1: Anatomical main group (1 letter)

Level 2: Therapeutic main group (2 digits)

Level 3: Pharmacological main group (1 letter)

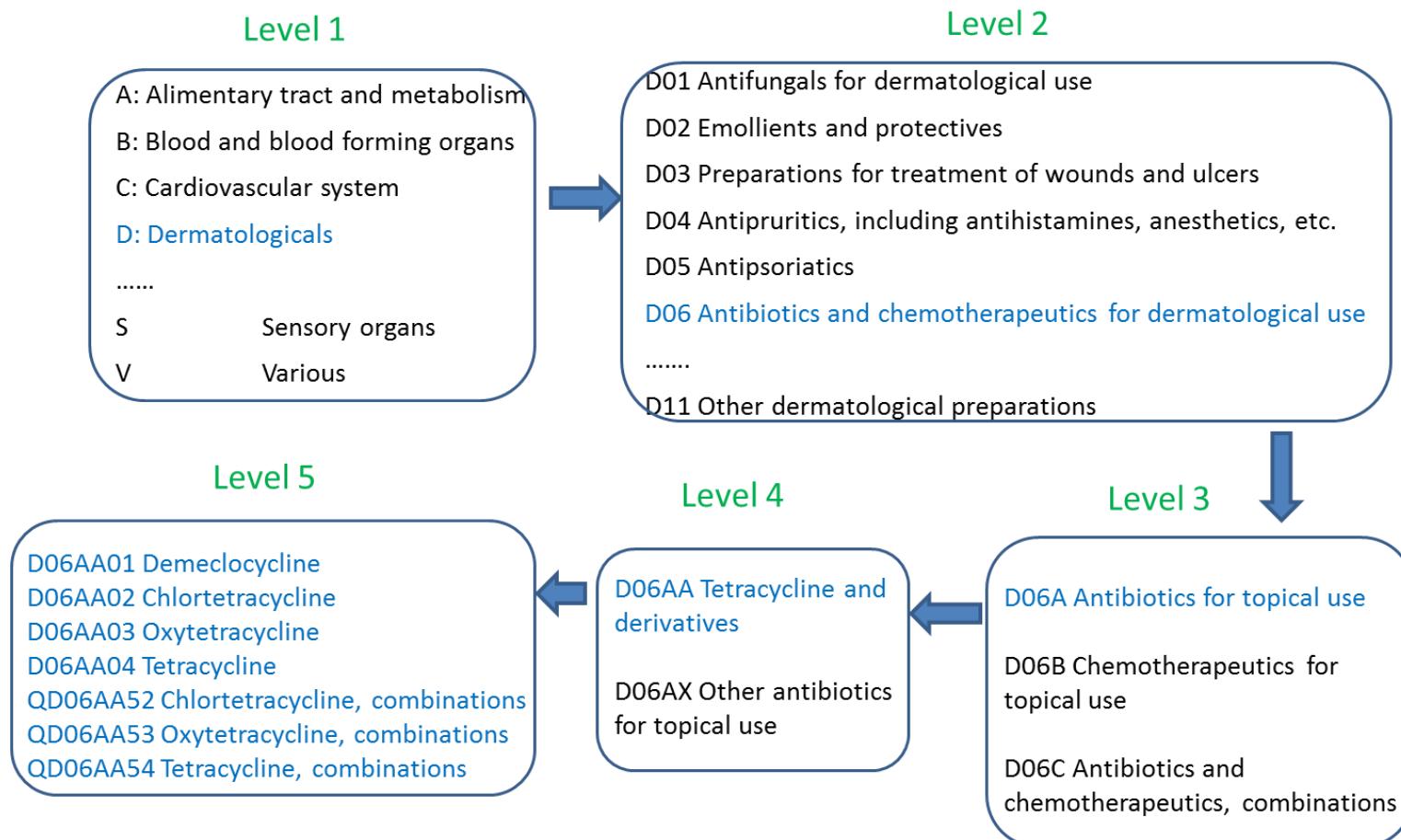
Level 4: Chemical main group (1 letter)

Level 5: Active substance main group (2 digits)

RESOURCES

(2) Integrated existing semantic domain resources

Example of ATC Structure



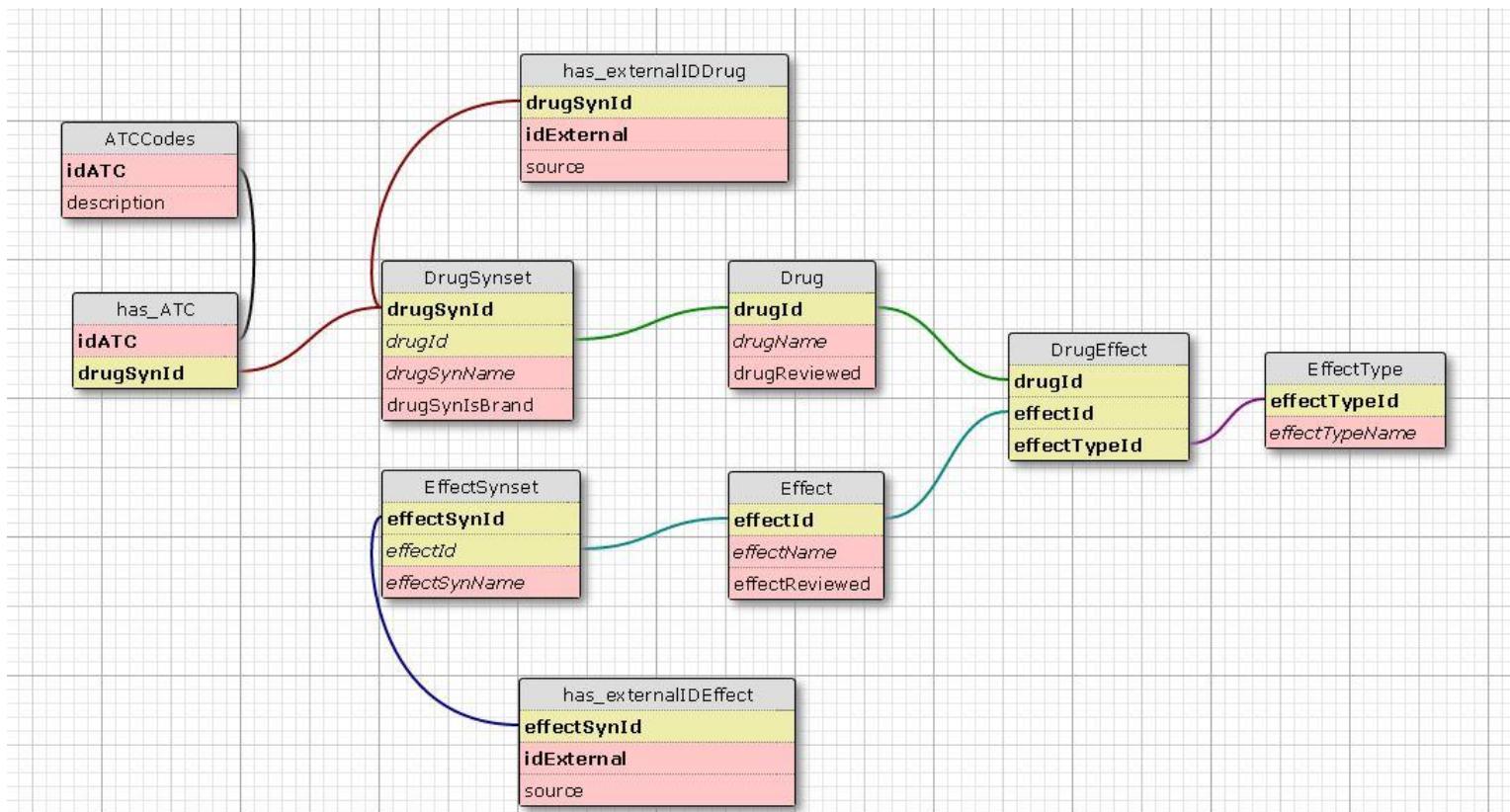
RESOURCES

(3) Integrated new semantic resources



63.000 relations

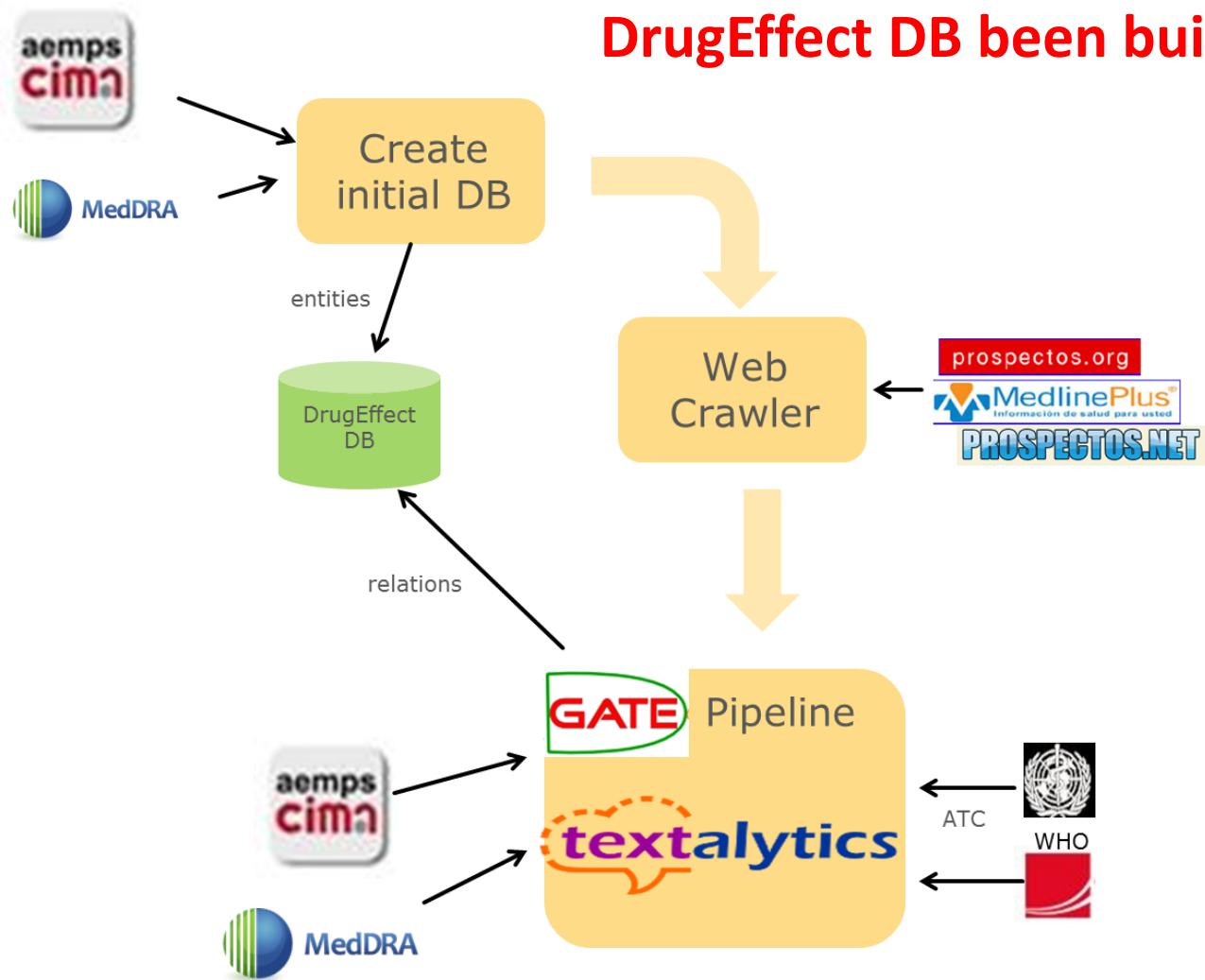
Spanish DrugEffect DB containing relations among drugs and effects



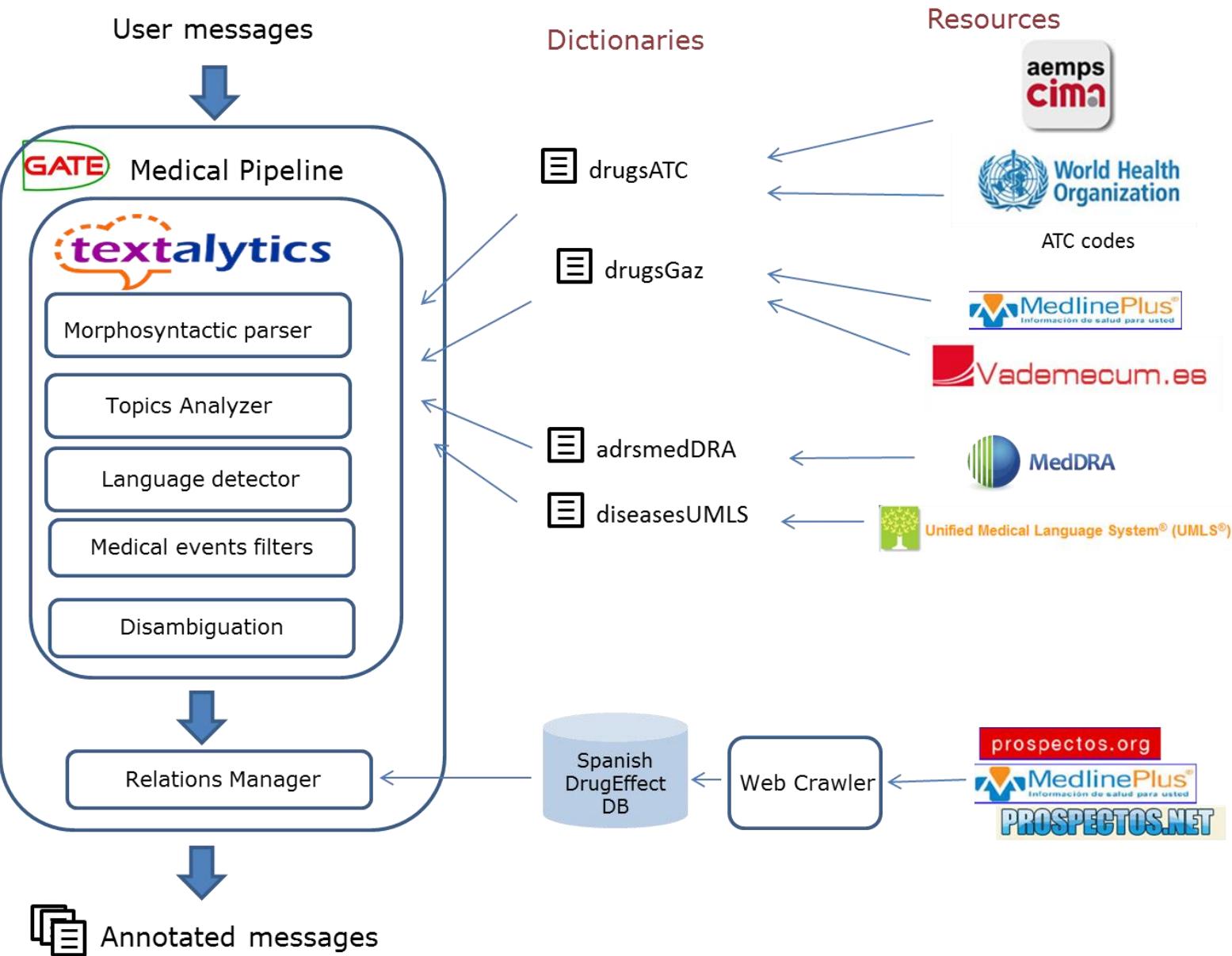
RESOURCES

(3) Integrated new semantic resources

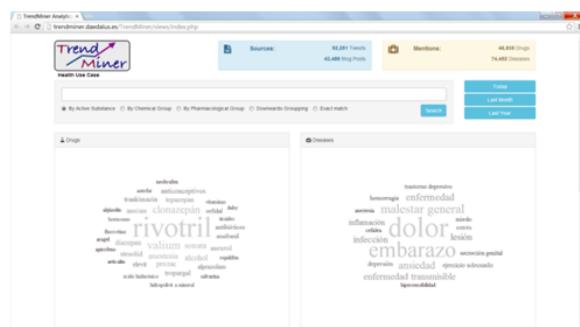
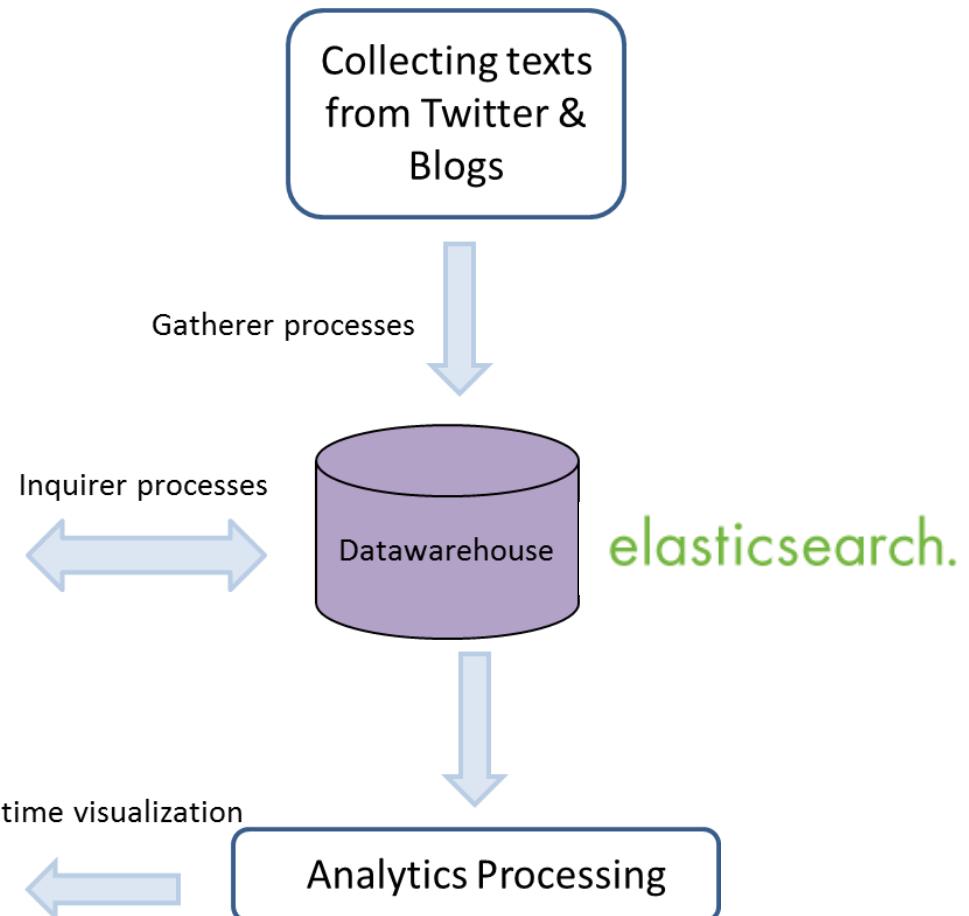
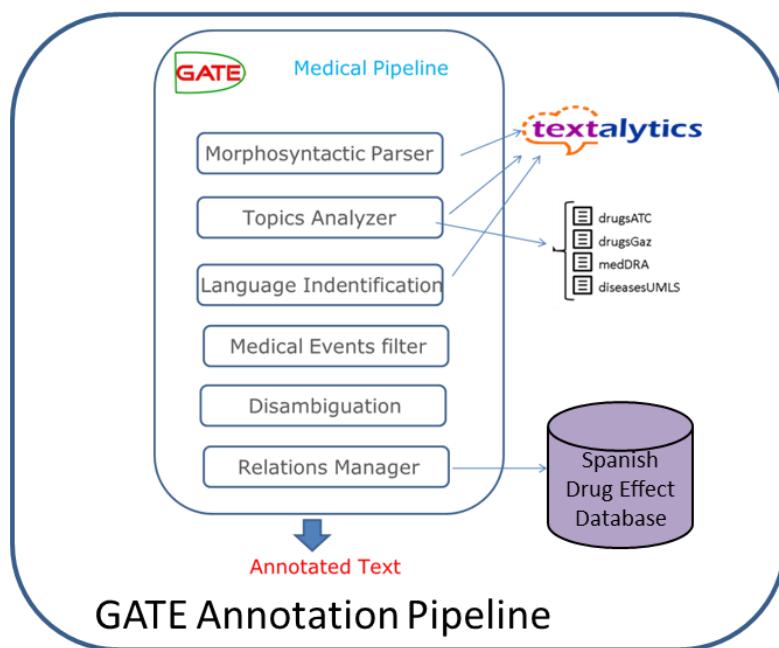
How has the Spanish DrugEffect DB been built?



LINGUISTIC PROCESSOR



REAL-TIME PROTOTYPE



Health monitoring Dashboard

REAL-TIME PROTOTYPE

- GATE Annotation Pipeline: Plugging developed for Textalytics
- Datawarehouse:
 - Implemented on elasticsearch with an ATC-based index structure:
January 19, 2015 , **2,401,613 Tweets** and **41,985 Blog Posts** annotated and indexed
 - Amazon infrastructure for processing and storing
- Searching:
 - Different search modes depending of ATC level or exact matching.
 - Obtaining and distinguishing indications, adverse effects and possible relations among drugs and effects.
 - Searching for co-occurrences (drug-disease, drug-drug,). Also machine learning as distant supervision approach (not in prototype)
- Visualization:
 - Timeline to view evolution of mentions with different granularity.
 - Viewing the annotated source text (tweets and posts)

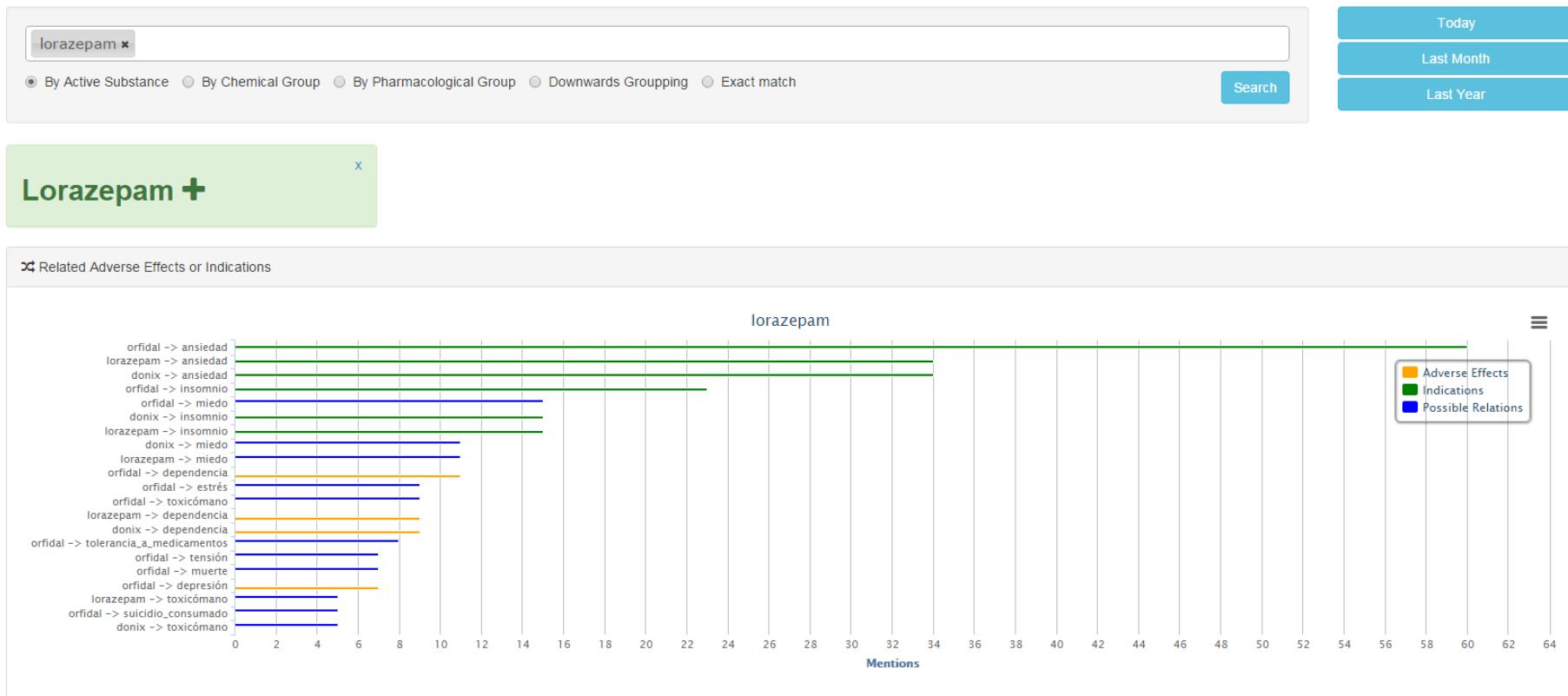
REAL-TIME PROTOTYPE



Health Use Case

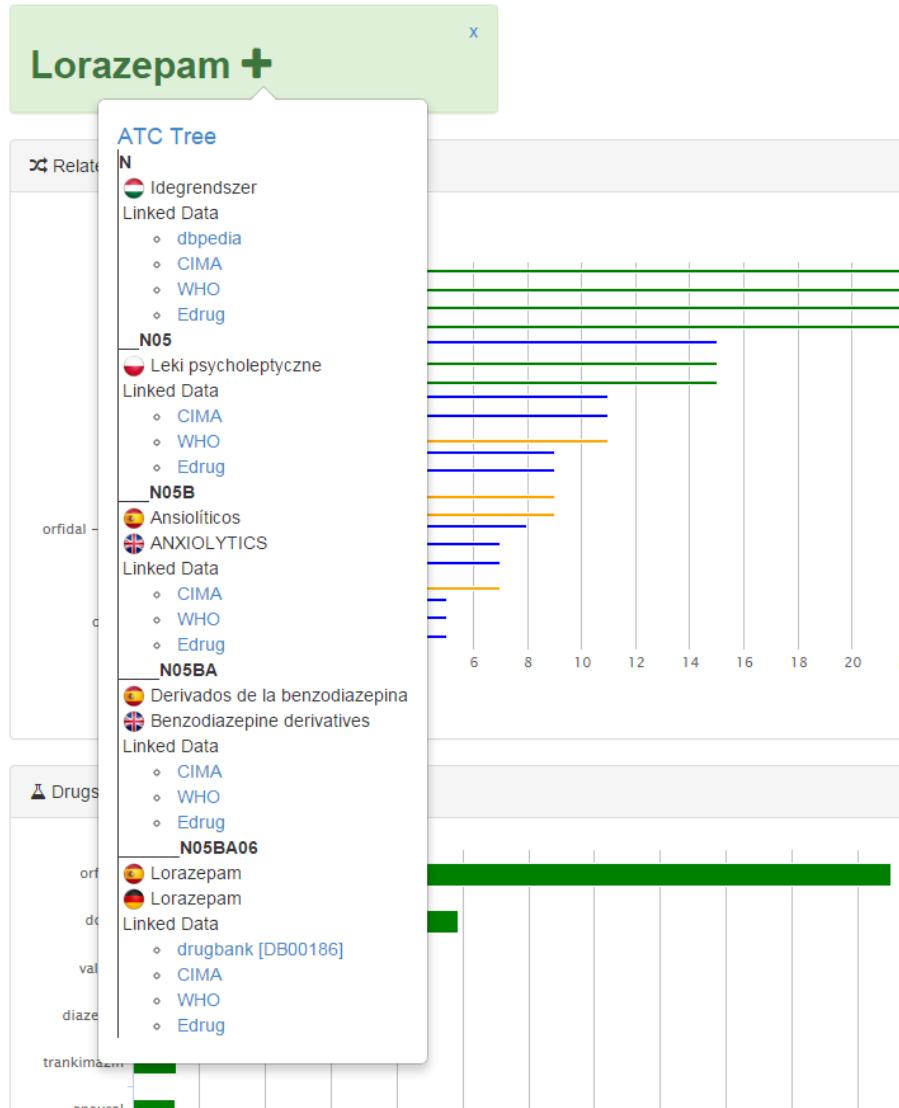
Sources:	1,110 Tweets 40 Blog Posts
----------	-------------------------------

Mentions:	4,566 Drugs 319 Medical Events
-----------	-----------------------------------



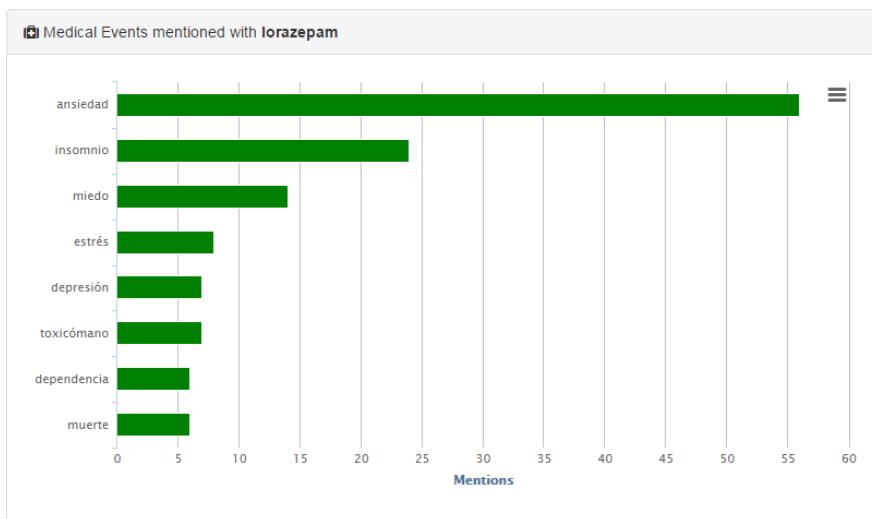
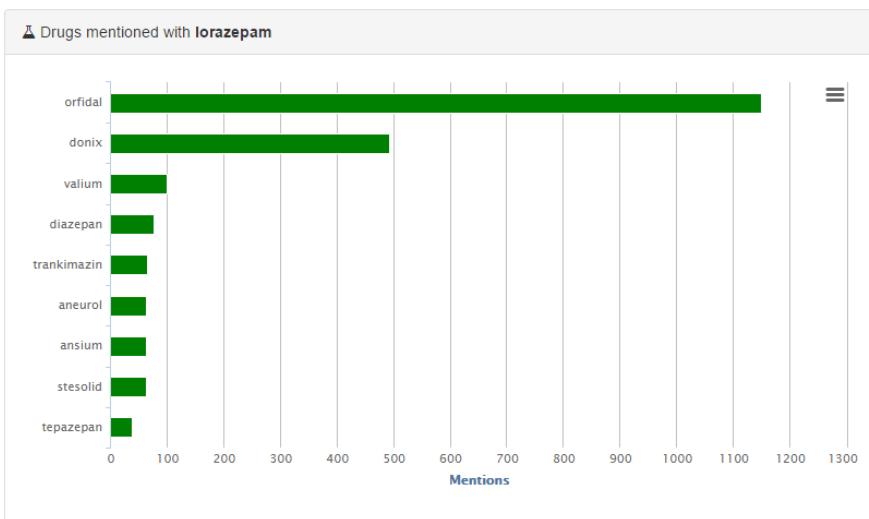
Aggregated data about effects related to drug lorazepam

REAL-TIME PROTOTYPE



Lorazepam ATC tree cross-lingual

REAL-TIME PROTOTYPE



Source Texts

 **macmartac** [2015-01-18T13:50:27 GMT]
A mí sólo se me contagian los **bostezos** si me tomo un **orfida** laconfesiondelas14h

orfidal
bostezos
☒ orfidal -> bostezos [possible relationship]

Drug-drug and drug-diseases co-occurrences for lorazepam active substance

REAL-TIME PROTOTYPE



Health Use Case



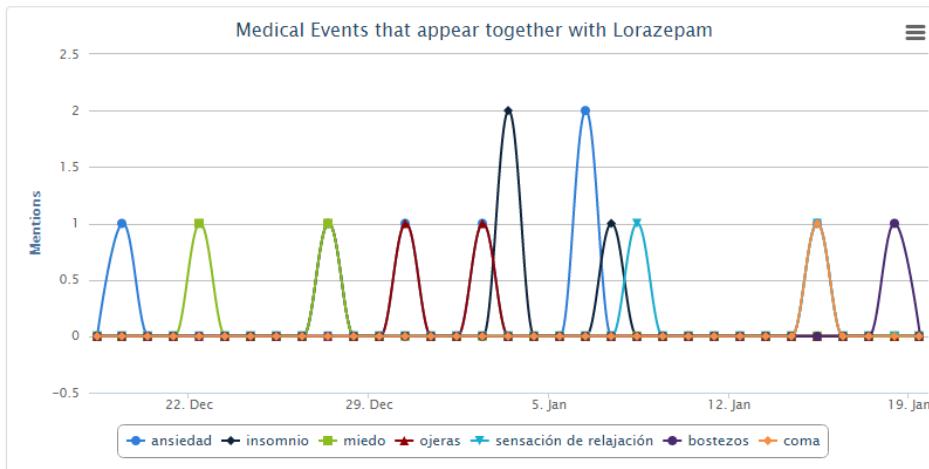
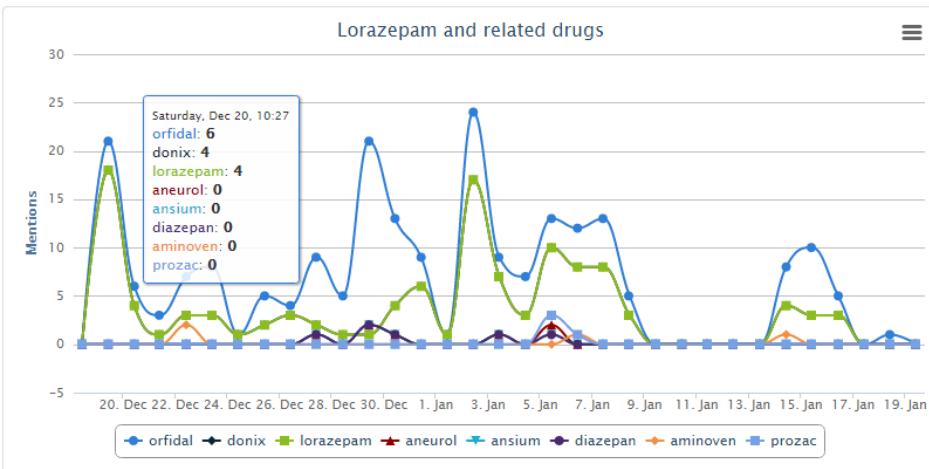
lorazepam x

By Active Substance By Chemical Group By Pharmacological Group Downwards Grouping Exact match

Search

Today
Last Month
Last Year

Lorazepam +



Timeline with evolution of *lorazepam* related drugs

ANNOTATION PIPELINE EVALUATION

- Using SpanishADR corpus (400 annotated comments from Forumclínic)

Drugs

Drugs	R	P	F-Measure
strict	0,68	0,85	0,76
lenient	0,68	0,85	0,76

- To enhance recognizing of misspelled drugs
- Include abbreviations for drug families
- Solve several ambiguities (alcohol, oxygen)

ANNOTATION PIPELINE EVALUATION

Effects

Effects	R	P	F-Measure
strict	0,43	0,75	0,54
lenient	0,47	0,83	0,6

- Low performance:
 - because colloquial expressions to describe an effect: *me deja KO* (it makes me KO) or *me cuesta más levantarme* (it's harder for me to wake up).
 - different lexical variations and abbreviations of the same effect.

ANNOTATION PIPELINE EVALUATION

Drug-effect relations

Window size		SpanishDrugEffectDB			Coocurrences		
		R	P	F-Measure	R	P	F-Measure
30	strict	0,08	0,57	0,14	0,63	0,44	0,52
	lenient	0,13	0,96	0,24	0,88	0,61	0,72
100	strict	0,1	0,34	0,16	0,74	0,26	0,38
	lenient	0,23	0,74	0,35	0,99	0,34	0,51
250	strict	0,12	0,32	0,17	0,17	0,75	0,33
	lenient	0,24	0,67	0,36	1	0,29	0,45

- Performed from annotated drugs and effects
- Low recall using SpanishDrugEffect DB because low coverage of effects, the lack of co-reference resolution and size of corpus (only 164 relations)
- Machine Learning tryed: distant supervision approaches

OTHER METHODS TO EXTRACT DRUG-EFFECT RELATIONS

1. Distant-supervision method using the database on a collection of 84,000 messages in order to extract the relations between drugs and their effects (instances of DB are positive examples)



Chronic diseases
Aprox 8.000 registered patients
Over 6 million of view pages



Total: 84,090 posts

- **Schizophrenia:** 26,234 (31.20%)
- **Depression:** 22,938 (27.28%)
- **Breast cancer:** 15,675 (18.64%)
- **Bipolar disease:** 12,573 (14.95%)
- COPD (Chronic obstructive pulmonary disease): 1,853 (2.20%)
- Ischaemic heart disease : 1,840 (2.19%)
- HIV/AIDS: 1,359 (1.61%)
- Obesity: 706 (0.84%)
- Take care: 419 (0.50%)
- Joint disease and arthritis: 276 (0.33%)
- Diabetes: 202 (0.24%)
- Colon cancer: 15 (0.02%)

OTHER METHODS TO EXTRACT DRUG-EFFECT RELATIONS

2. To classify the relation instances, we used a kernel method based only on shallow linguistic information of the sentences.
3. Regarding Relation Extraction of drugs and their effects, the distant supervision approach achieved a recall of 0.59 and a precision of 0.48

POSSIBLE EXTENSIONS

- Development of cross-lingual approach for ATC codes integrating ATC ontology in collaboration with DFKI
- Populate Health TM ontologies from semantic resources in collaboration with DFKI
- Many possibilities of customization:
 - Tracking specific drugs appearing with different diseases or effects
 - Information Extraction from unstructured data (i.e., detection of allergies in EHRs)
 - Helping on-line health blogs and forums managers

References

Aplicación Distant supervision a la extracción de relaciones

Isabel Segura-Bedmar, Paloma Martínez, Ricardo Revert , Julián Moreno-Schneider, (2015). Exploring Spanish Health Social Media for detecting drug effects, BMC Medical Informatics and Decision Systems, January, 2015, Volumen: In press.

Demos

Santiago Peña-González, Isabel Segura-Bedmar, Paloma Martínez, José Luis Martínez Fernández,(2014). ADRSpanishTool: a tool for extracting adverse drug reactions and indications, September, 2014, Procesamiento del Lenguaje Natural, Volumen: 53, Páginas: 177-180, [url](#).

Corpus para entrenamiento y test

María Herrero Zazo, Isabel Segura-Bedmar, Paloma Martínez, Thierry Declerck, (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions, Journal of Biomedical Informatics (IF 2012: 2.131), October, 2013, Volumen: 46, Número: 5 DOI: 10.1016/j.jbi.2013.07.011, Páginas: 914-920, [url](#).

References

Corpus ADRs y enfoques basados en diccionarios

Isabel Segura-Bedmar, Ricardo Revert , Paloma Martínez, (2014). Detecting drugs and adverse events from Spanish social media streams, Gothenburg, Sweden, April, 2014, Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi), Association for Computational Linguistics, Páginas: 106-115, [pdf](#).

Isabel Segura-Bedmar, Santiago Peña-González, Paloma Martínez, (2014). Extracting drug indications and adverse drug reactions from Spanish health social media, Proceedings of the BioNLP 2014 workshop, USA, June, 2014, Association for Computational Linguistics, Páginas: 98-106, [pdf](#).

Tarea Semeval DDIExtraction 2013 (<http://www.cs.york.ac.uk/semeval-2013/task9/>)

Isabel Segura-Bedmar, Paloma Martínez, María Herrero Zazo, (2014). Lessons learnt from the DDIExtraction-2013 shared task, January, 2014, Journal of Biomedical Informatics (IF 2012: 2.131), Elsevier, ISSN: 1532-0464, Volumen: 51, Páginas: 152-164, [url](#).



Universidad
Carlos III de Madrid



Contacto: Paloma Martínez

E-mail: pmf@inf.uc3m.es

@Grupo_LaBDA
labda.inf.uc3m.es