



Abstract

The Spanish Agency of Medicines and Medical Devices (AEMPS), has created the Public Pharmacoepidemiological Information Database (BIFAP), from the aggregation of the different regional healthcare databases.

Due to the heterogeneity of the terminologies used in the distinct, regions of the country, it has been decided to standardize all the descriptors of health problems in free text to SNOMED-CT.

In addition, the use of SNOMED-CT in the coding of health records allows the creation of RefSets and correspondence between different coding systems such as ICD10 and ICPC2.

We have applied simple natural language processing techniques and have carried out an approximation of the similarity of the descriptors registered in BIFAP to the different synonyms in the Spanish edition of SNOMED-CT, obtaining coverage of ~80% of the records of health problems in BIFAP, corresponding to a small percentage of all the different forms in the text that we find in this field.

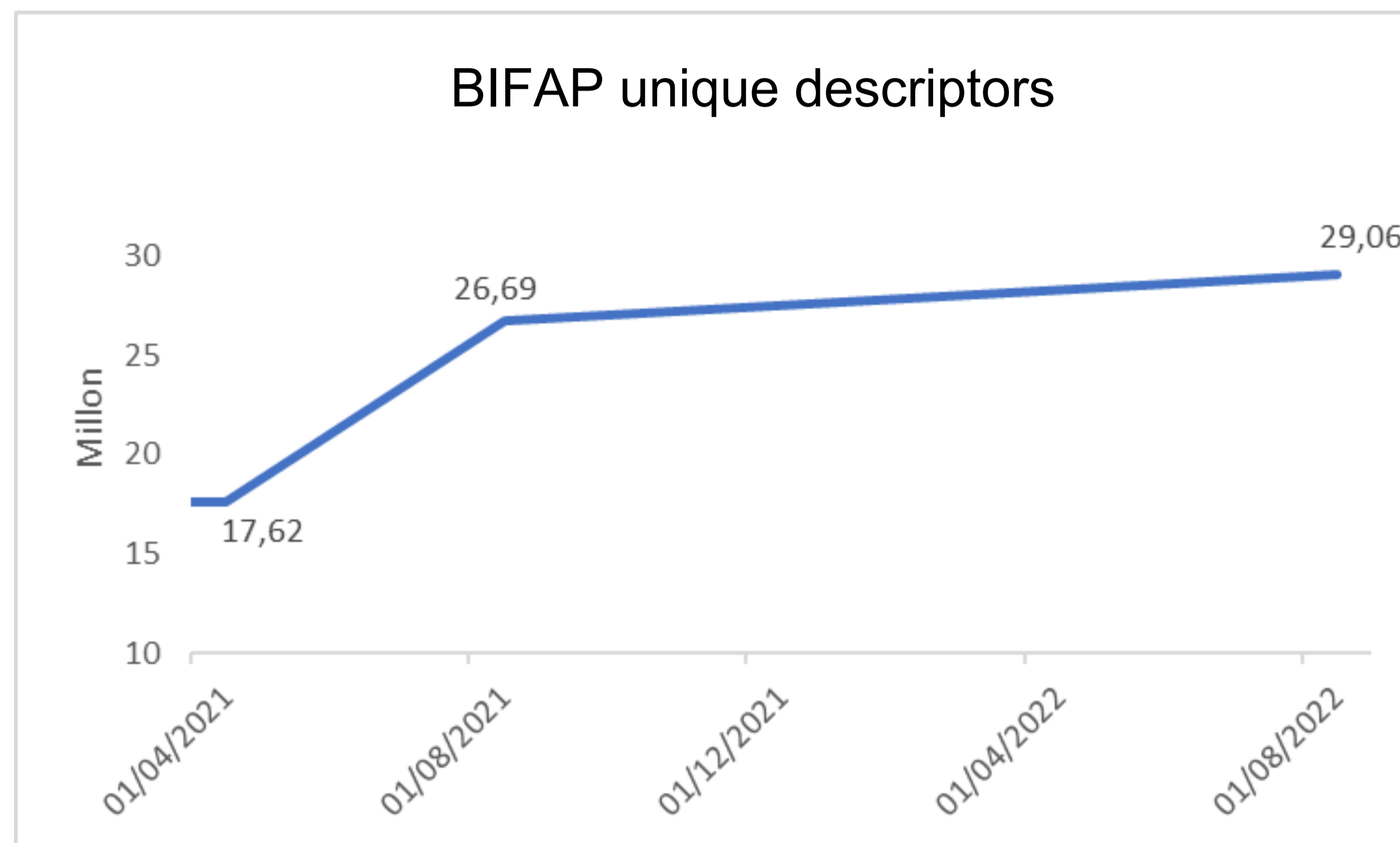
Topic

- SNOMED-CT uses of normalization and integration tool from different terminologies.
- Use of tools for mapping suggestions management.
- Use of tools for automatic mapping suggestions.
- Keypoints in reviewing mapped medical concepts.
- What I would have liked to know before starting a project like this?





BIFAP Diagnostic descriptors



Free text fields in database are growing without improving accessibility to normalized data.

Previous attempt to normalize free text diagnostics to a ICP2-derived terminology was unsuccessful, due the limited size of labels.

BIFAP Initial status

	ICPC2			
	Event count	%	Unique descriptors	%
Normalized	346.176.314	68,7%	285.405	1,6%
Unnormalized	157.771.630	31,3%	17.443.442	98,4%
total	503.947.944		17.728.847	

Mapping strategies

Mapping pipeline:

- Exact literal matching
- Levenshtein/Damerau-Levenshtein distances
- NLP techniques





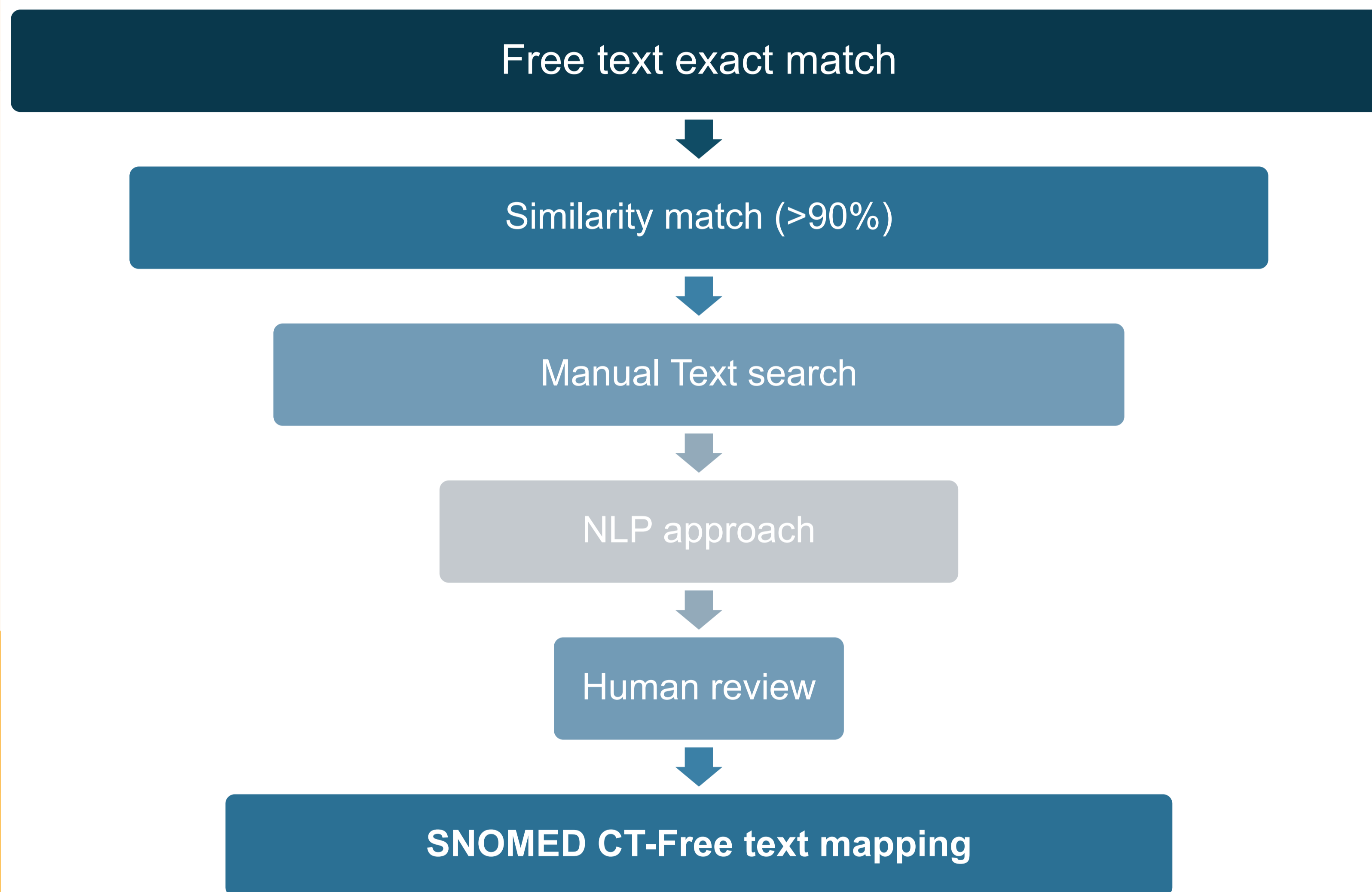
Introduction

BIFAP first mapping approach

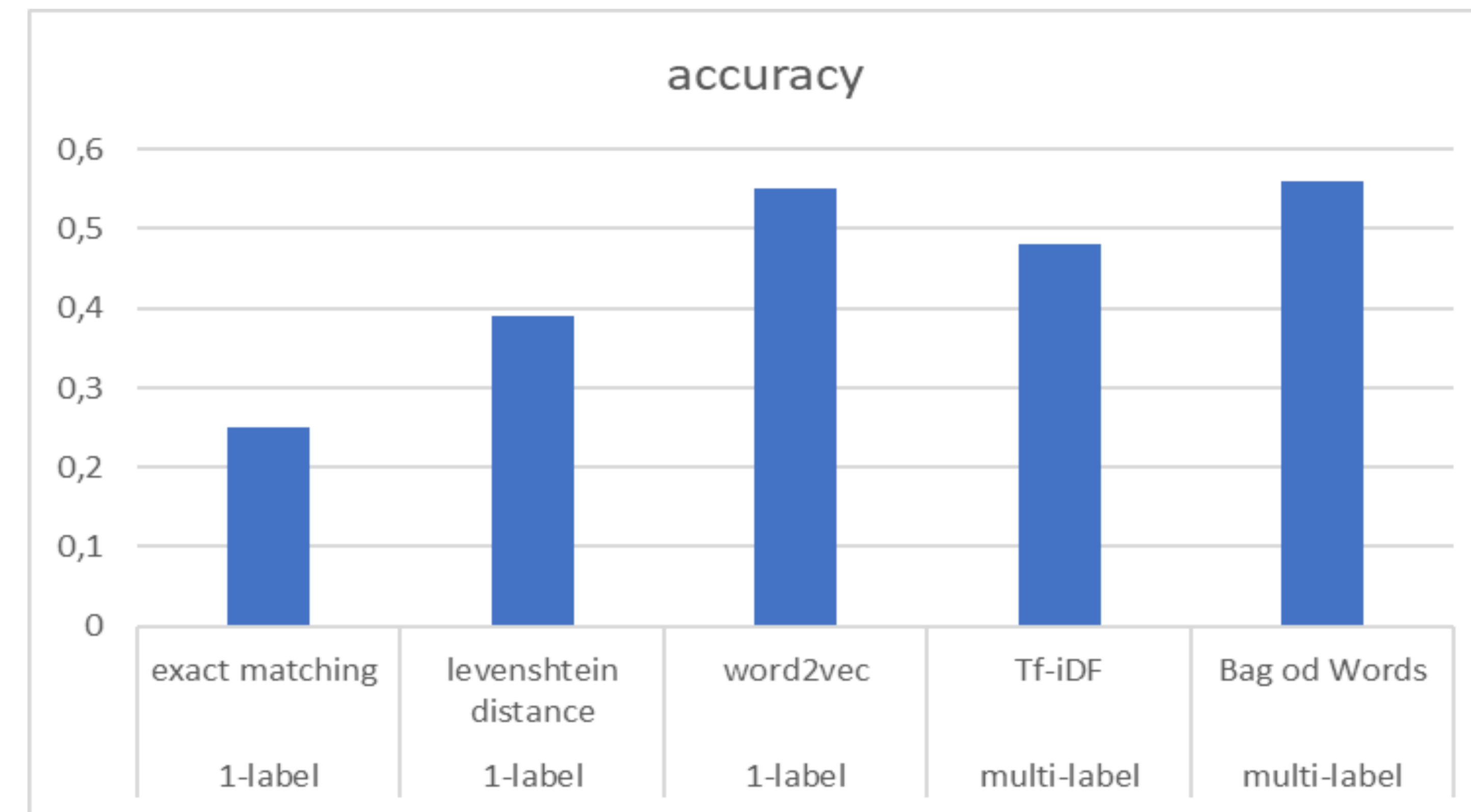
Methods

Results

Discussion



BIFAP NLP techniques metrics



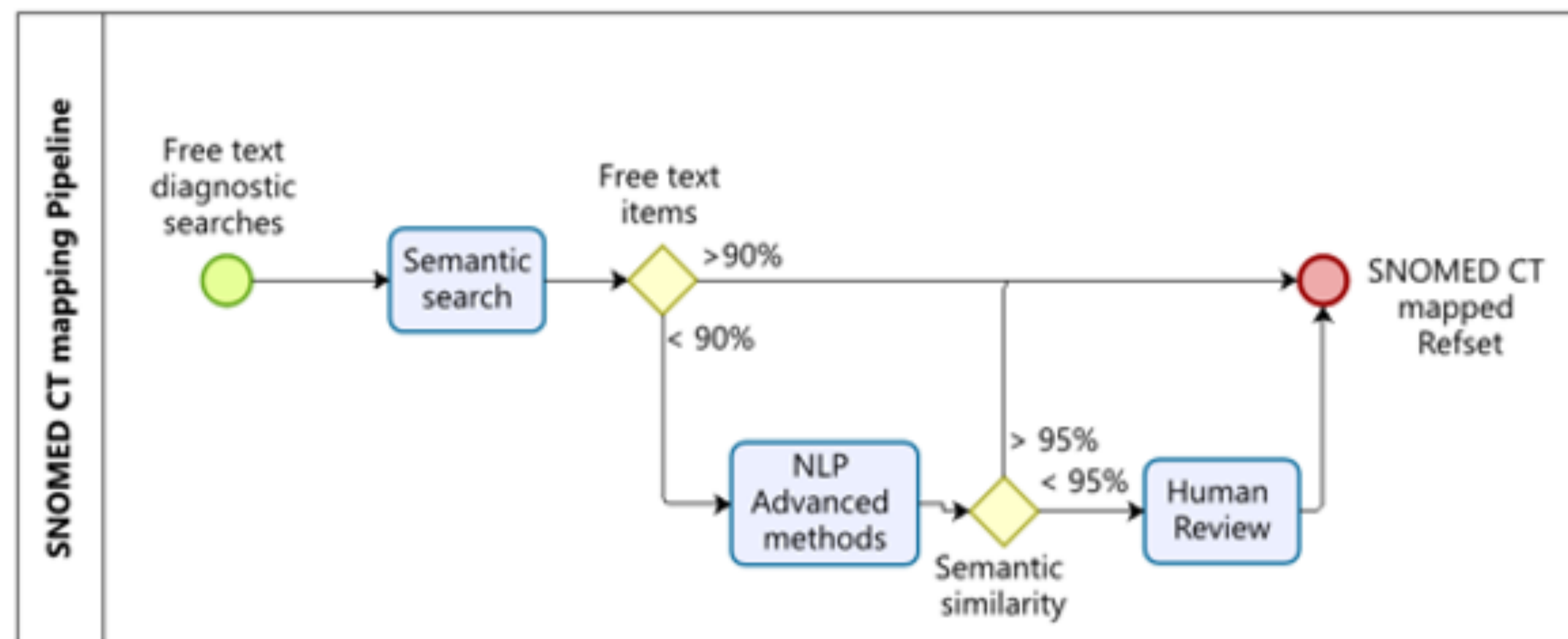
BIFAP up-to-day-results

	ICPC2				SNOMED CT			
	Event count	%	Unique descriptors	%	Event count	%	Unique descriptors	%
Normalized	472.031.050	65,9%	280.690	0,9%	586.774.977	81,9%	813.931	2,8%
Unnormalized	244.269.767	34,1%	29.285.908	99,1%	129.525.840	18,1%	28.752.667	97,2%
total	716.300.817		29.566.598		716.300.817		29.566.598	



Conclusions

- Combining different methods improve the mapping tasks



- Once the effectiveness of the different approaches was verified, it was decided not to lose the Human review, to create a pipeline where these techniques were incorporated (in the form of an ensemble) according to the refsets that were needed in the studies.
- Semantic searches are made into the database, they are suggested to the algorithm that maps them and they go to the Normalization Unit to confirm and change those that have not been correctly assigned.

What I would have liked to know before starting a project like this?

- Involve all kind of roles in the project from beginning
- Combine different strategies by steps (no need all to start)
- Take small steps and reuse them
- Add new methods in your toolbox
- Find bottlenecks quick to resolve before they grow
- Keep updated

Future Directions

- Improving model to semiautomatics mapping
- Use NLP new techniques like transformers pipelines
- Implement SNOMED CT shared tools to end users
- Create internal RefSet repositories with automatic updates

1. References

- [BIFAP Base de datos para la Investigación Farmacoepidemiológica en Atención Primaria \(aemps.es\)](http://aemps.es)
- Gonzalo Navarro. 2001. A guided tour to approximate string matching. ACM Comput. Surv. 33, 1 (March 2001), 31–88. <https://doi.org/10.1145/375360.375365>
- CHURCH, K. (2017). Word2Vec. Natural Language Engineering, 23(1), 155-162. doi:10.1017/S1351324916000334
- Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. ArXiv, abs/1706.03762.D SNOMED CT implementation course, mapping strategies.
- SNOMED CT Editorial guide. <http://snomed.org/eg>