Is it a lay or medico-scientific concept? Proposals for an automatic classification

Paulo Miguel Santos and Carla Teixeira Lopes
Department of Informatics Engineering, Faculty of Engineering, University of Porto, Porto, Portugal
INESC TEC, Porto, Portugal
{up201403745, ctl}@fe.up.pt

INTRODUCTION

Systems that suggest queries in lay and medico-scientific terminology improve **health information retrieval** by who is not a health professional.

This leads to the need for **automatic recognition** if **concepts** are **lay or medico-scientific, in order to** provide more focused support and better retrieval services to users.

Several approaches are proposed to compute the degree of association of a concept to lay and medico-scientific terminology.

using

Different vocabularies

Cosine similarity to measure the closeness of concepts with subsets of those thesauri.

METHODOLOGY

Name	Used Vocabulary	Number of expressions	Classification method	Cosine similarity computation
CHV-single	CHV	146,334	Threshold	Single document
CHVpref-single	CHV-preferred	57,796		
UMLSpref-single	UMLS-preferred	30,060		
MedPlus-single	Medline Plus	3,329		
CHVUMLSpref-single	CHV-preferred and UMLS-preferred	87,856	Max(simCHV, simUMLS)	
CHV-multi	CHV	146,334	Threshold	Several documents
CHVpref-multi	CHV-preferred	57,796		
UMLSpref-multi	UMLS-preferred	30,060		
UMLSpref-multi	Medline Plus	3,329		
CHVUMLSpref-multi	CHV-preferred and UMLS-preferred	87,856	Max(simCHV, simUMLS)	

Single document

$$sim\begin{pmatrix} medical\ concept, \\ vocabulary \end{pmatrix} = cos\ \theta$$

The cosine similarity is calculated between the original medical expression and the document containing all the expressions of the considered vocabulary.

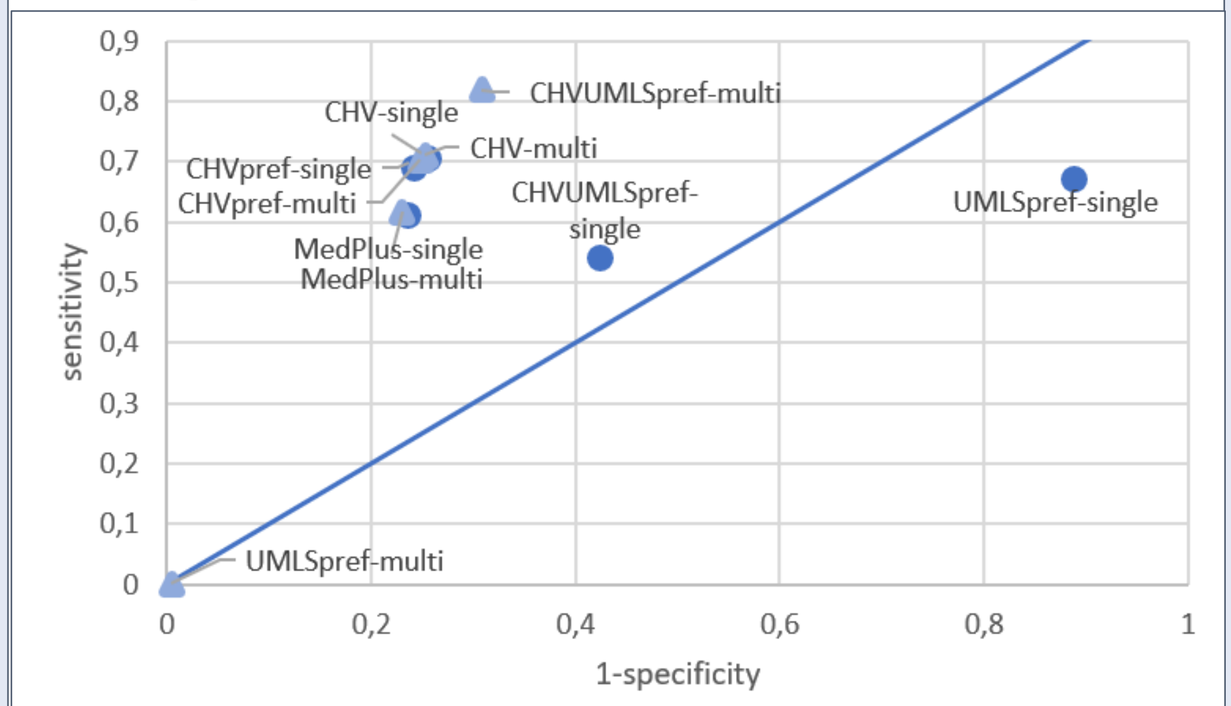
Several documents

$$\frac{\sum_{i=1}^{N} sim\begin{pmatrix} medical\ concept, \\ vocabulary\ expression \end{pmatrix} = cos\ \theta}{\sum_{i=1}^{N} vocabulary\ expression}$$

Each expression is a different document and the similarity is computed with all of them. In the end we average these similarities.

EVALUATION

We used the Multilingual Glossary of technical and popular medical terms as ground truth.



Method	Threshold	ACC	SEN	SPC	ROCD
CHV-single	0,0017	0,7288	0,7142	0,7500	0,3797
CHVpref-single	0,0017	0,7249	0,6984	0,7635	0,3833
UMLSpref-single	0,0005	0,4520	0,6814	0,1178	0,9379
MedPlus-single	0,0006	0,6821	0,6219	**0,7699**	0,4426
CHVUMLSpref-single	-	0,5638	0,5519	0,5812	0,6133
CHV-multi	0,0001	0,7255	0,7104	0,7476	0,3842
CHVpref-multi	0,0001	0,7226	0,7027	0,7516	0,3874
UMLSpref-multi	0,0044	0,4054	0,0005	0,9952	0,9995
MedPlus-multi	0,0001	0,6792	0,6169	**0,7699**	0,4469
CHVUMLSpref-multi	-	**0,7673**	**0,8191**	0,6918	**0,3572**

CONCLUSIONS

CHVUMLSpref-multi performs better than CHVUMLSpref-single. In this case, it is better to consider each expression a separate document.

Other methods have a similar performance regardless of how the cosine similarity is computed, with the exception of the UMLSpref approach.

The methods that use **CHV** outperformed most of the other methods.

The best method uses CHV-preferred and UMLS-preferred expressions as vocabularies and chooses the maximum value of similarity with their expressions.

ACKNOWLEDGEMENTS

Work supported by: **Integrated project NanoSTIMA** Project "NORTE-01-0145-FEDER-000016" is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).