Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 129-136 recibido 01-04-2019 revisado 08-05-2019 aceptado 15-05-2019

Landscaping Language Technologies using Topic Modeling and Graph Analysis: Overview of the Spanish Contribution

Caracterización del sector de Tecnologías del Lenguaje mediante modelado de tópicos y análisis de grafos: Visión general de la participación española

Doaa Samy[1,3], David Pérez-Fernández[1], Jerónimo Arenas-García[1,2]
[1]Secretaría de Estado de Avance Digital, Spain
[2]Universidad Carlos III de Madrid, Spain
[3]Instituto Ing. del Conocimiento (IIC), Madrid, Spain and Cairo University, Egypt
doaa.samy@iic.uam.es, dperezf@mineco.es, jeronimo.arenas@uc3m.es

Abstract: This paper aims at landscaping the Human Language Technologies (HLT) sector by applying topic modeling and graph analysis to study the scientific literature in ACL Anthology with special emphasis on the Spanish participation. The analysis takes into account the structured and unstructured data to offer an overview of the HLT landscape in Spain identifying main underlying themes and its evolution in the last years compared to the international HLT community. Results obtained are represented through an interactive visualization to allow the exploration of the HLT landscape in the time frame 1983-2018.

Keywords: Human Language Technologies, Topic Modeling, Latent Dirichlet Allocation (LDA), Louvain modularity algorithm, Natural Language Processing

Resumen: El presente trabajo aplica herramientas de modelado de tópicos y análisis de grafos para caracterizar el sector de Tecnologías del Lenguaje (TL) en España. Para ello, se estudian el repositorio de ACL Anthology. Este análisis tiene en cuenta los datos estructurados y no-estructurados en dichas fuentes con el fin de retratar el panorama actual en términos de temáticas subyacentes y su evolución en los últimos años en comparación con la comunidad internacional. Los resultados se presentan mediante una visualización interactiva que permite navegar en el espacio de TL en el intervalo temporal 1983-2018.

Palabras clave: Tecnologías del Lenguaje, Modelado de Tópicos, Latent Dirichlet Allocation (LDA), Algoritmo de modularidad de Louvain, Procesamiento del Lenguaje Natural

1 *Introduction*

The amount of data available from scientific production is increasing rapidly. Dynamic innovative tools are needed to analyze this data in its structured and unstructured formats to offer better insights on the situation and the progress in different fields. Bibliometrics and Scientometrics approaches have opted for statistical approaches and indicators based on structured data (Xu et al., 2017) (Serenko et al., 2010) (Langhe, 2016) (Clauset, Newman, and Moore, 2004). However, Natural Language Processing (NLP) and Artificial Intelligence can offer a complementary approach by discovering the underlying semantics in the unstructured text.

In this paper, we opt for a comprehensive hybrid approach which has been previously adopted in Corpus Viewer, a tool developed by the authors that analyzes the R&D&i space in general, including scientific publications, funded projects, patents, etc.

Building on the above previous experience, we address the HLT field as a rapidly growing and interdisciplinary domain which would rather benefit from this approach. Moreover, HLT has recently been in the centre of AI advances as an enabling technology

ISSN 1135-5948. DOI 10.26342/2019-63-14

in different fields: Human Computer Interactions, Internet of Things, Smart Solutions, etc. This is reflected in an increasing demand by the different sectors to include NLP components in their solutions. Also, it is reflected in several areas of applications and research lines.

A key starting point for a strategic plan to promote HLT in Spain and eventually in Hispanic countries, is: 1) to understand and characterize the sector and its dynamics in general and 2) to characterize the Spanish sector. In this line, the present study stems out from the following basic questions:

- What are the main thematic lines addressed by HLT and how these lines are reflected in the scientific literature?
- How these thematic lines have evolved along the years?
- What is the participation of Spanish key-players and their contribution within the general HLT landscape?

Answers to these questions would provide insights for the current strategic plan in Spain and would add value to the sector for a better self-positioning within the international landscape.

Although our approach is completely general and could be used with other datasets, to landscape the HLT sector we rely primarily on the ACL Anthology[1](Bird et al., 2008), a repository of computational linguistics and NLP papers, including open access to the full text in pdf format. In addition to this, we used Semantic Scholar[2] to enrich the information from ACL adding abstracts, entities and citations. The ACL Anthology is a valuable resource that has been used in a number of studies for example (Hall, Jurafsky, and Manning, 2008) applied topic modeling to study the history of ideas in the ACL Anthology. The ACL Searchbench was developed in 2011[3] (Schäfer et al., 2011) allowing the search by keywords, full text, affiliation, etc. However, our approach is completely different since it pretends to landscape the information space of the ACL Anthology semantically and dynamically in time. Others enriched the Anthology with semantic (Gábor et al., 2016) or co-reference annotation (Schäfer, Spurk, and Steffen, 2012). More studies included scientific term mining (Jin et al., 2013), studying gender aspects (Vogel and Jurafsky, 2012), etc.

The rest of the paper is divided into 3 sections. The methodology is described in Section 2 underlining the data sources and the experiments. Results and conclusions driven are discussed in Sections 3 and 4.

2 Methodology

Our methodology consists of the following stages, that will be described in the next subsections:

1. Identifying relevant datasets representing HLT scientific literature and extracting the metadata fields relevant to the study.
2. Selecting the data subset representing the Spanish participation.
3. Processing the unstructured text using basic tokenization and segmentation through an NLP Pipeline. The final output of this stage is the text lemmatized and a set of n-grams.
4. Topic modeling.
5. Construction of semantic graphs based on the interdistance between the topic vectors identified in the previous step.
6. Semantic community identification and characterization.
7. Visualization of the results using a graph visualization tool, such as Gephi (Bastian, Heymann, and Jacomy, 2009).

2.1 Enriching ACL Anthology metadata and identifying Spanish contributions

Two main datasets are used in this study: 1) ACL Anthology as a representative dataset of HLT scientific literature and 2) Semantic Scholar as an additional resource used for enriching the dataset with extra information which is not available in ACL Anthology. In Semantic Scholar we have access to the following metadata that is not available in ACL: abstract, citations and "entities". It is important to point out that the term "entities" applied by Semantic Scholar is not necessarily *named entities*, but rather candidates of

[1] https://www.aclweb.org/anthology/

[2] https://www.semanticscholar.org/

[3] http://aclasb.dfki.de/#

domain specific terms or n-grams in the abstract.

As of March, 2019, ACL Anthology hosts 48, 286 entries. The whole website was downloaded using a web crawling tool, including paper, author, and event metadata. A total of 47, 198 full papers in pdf format where also crawled from the web. In order to enrich the information with metadata available in Semantic Scholar, we carried out a matching procedure, so that for every paper in the ACL collection we searched for an appropriate match in Semantic Scholar. To avoid wrong assignations, the process was carried out ensuring that matched papers were published in the same year, and that at least one of the following conditions was satisfied:

- DOI match: matched papers should have identical DOIs.
- Title match: matched papers should have identical lowercase titles.
- URL match: the URL from which the ACL paper was retrieved is listed in the URL field of the matched paper from Semantic Scholar.

As a result of this process, we obtained a dataset of 42, 396 papers. A simple inspection of the unmatched ACL papers revealed that most of them correspond to lists of authors, cover pages of conference proceedings, book reviews, etc.

In addition to the general ACL corpus, we created a subcorpus of ACL Anthology reflecting the Spanish HLT contribution in the field (to which we will refer in the following as ACL-SPA). Inclusion of papers in the ACL-SPA subset was based on the criteria that at least one of the authors is affiliated to a Spanish institution.

Author affiliation is not available in either ACL or Semantic Scholar. Though affiliations are available in ACL Searchbench, they can not be readily applied since institutions names are not normalized and it does not contain all papers currently available in ACL. Therefore, we used a simple approach for detecting author affiliation: using regular expressions for email detection. Based on detected emails, we filtered the following domains: '.es', '.cat', '.eus', and '.gal'. We selected also '.edu' domains of Spanish Universities. Finally, we also searched for appearances of the word 'Spain' in the first page of the paper followed by human check to make sure only papers with Spanish affiliations were selected. The ACL-SPA subset contains 1, 408 papers, representing barely 3.32% of the total number of papers in ACL.

2.2 Unstructured Text Preprocessing

Available text for all the papers in the ACL Anthology was processed to obtain a valid document representation for topic modeling. For the study, we used the following unstructured text information:

- The title of the paper
- The list of entitities provided by Semantic Scholar
- The abstract, also available from Semantic Scholar
- The text of the full paper extracted from the pdf files using the *tika-python* library[4], which is based on the popular Apache Tika toolkit[5].

In order to build the vocabulary and lemmatize the text, we used the librAIry NLP toolkit (Badenes-Olmedo, Redondo-Garcia, and Corcho, 2017) which provides the following NLP tools: 1) Part-of-Speech Tagger (we keep nouns, adjectives, verbs, and adverbs), 2) Lemmatizer, 3) N-Grams Identifier and 4) Wikipedia Resource Finder. In addition to this, we removed stopwords using a generic list of stopwords, as well as a list of typos and common terms in the field identified by the authors that provide little semantic content in this context (e.g., 'document' , 'word'). Finally, we also removed from the vocabulary any word that appeared in less than 7 documents or in more than 40% of them.

After text preprocessing, we obtained a vocabulary of 16, 075 entries, which increased to 97, 053 when using the full papers. In the latter, many typos are still present due to common errors in the pdf extraction.

2.3 Topic Modeling using LDA

To analyze the ACL document collection and extract its underlying topics in an automatic way, we recur to Latent Dirichlet Allocation (LDA), a well-known tool for probabilistic

[4] `https://github.com/chrismattmann/tika-python`

[5] `https://tika.apache.org/`

topic analysis first proposed in (Blei, N, and Jordan, 2003). Although later refinements of LDA have appeared in the literature, the original algorithm suffices our needs in the paper, and has the advantage of availability of very efficient implementations. In this paper, we use the Mallet implementation (McCallum, 2002), which is based on Collapsed Gibbs Sampling and supports multithreading for very fast and scalable execution.

In LDA, a topic is characterized by a probability distribution over the complete vocabulary. Its generative model assumes also that each document is generated as a mixture of the different topics, where the sum of the topic proportions for each document is one. Therefore, to train the LDA model we provide as input the bag of word representations of all documents in the corpus and the number of topics (K), and the output we obtain consists of:

1. The length-K vectors representing the topic proportions for each document, $\boldsymbol{\theta}_d$, $d = 1, \ldots, \#\text{docs}$.
2. The vectors representing the vocabulary distribution for each topic, $\boldsymbol{\beta}_w$, $w = 1, \ldots, \#\text{vocabsize}$.
3. The relative importance of each topic.

Selection of the number of topics is always subject to a tradeoff between granularity of the topics (with a larger number of topics we observe very specific topics that do not appear for very few topics) and appearance of garbage topics (more numerous when using a large number of topics).

2.4 Semantic Graph Construction and Community Detection

Topic models are a very useful tool for finding out relevant topics inside a collection of text documents. However, there are some questions that remain unanswered with an LDA model. For instance, if we would like to know the number of papers about Machine Translation, the information on vectors $\boldsymbol{\theta}_d$ can only give an approximation, since many different papers will have a non-zero contribution to the topic. In other words, in principle all papers can belong to all topics but with a variable degree.

In this paper, we follow an alternative approach consisting of the following two steps: 1) building a semantic graph of papers, in which two papers are connected if they are semantically similar, and 2) finding groups of connected papers in the graph.

Unlike co-citation graphs, in our analysis we aim to calculate semantic graphs using the topic representation for each paper. The underlying hypothesis for prefering semantic graphs is that co-citation graphs do not contain links for all pairs of semantically related papers. Besides, some of the links in co-citation graphs may also be between papers from different topics. In short, we expect the semantic graph to be more complete and less noisy than the corresponding co-citation graph. An additional advantage is that link weights are real in contrast to co-citation graphs links.

In order to measure the semantic similarity between two papers, we compute the Jensen-Shannon divergence between their topic vectors,

$$\mathrm{JS}(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}) = \frac{1}{2} \sum_i \frac{\theta_d^{(i)} \log 2\theta_{d'}^{(i)} + \theta_{d'}^{(i)} \log 2\theta_d^{(i)}}{\theta_d^{(i)} + \theta_{d'}^{(i)}}, \tag{1}$$

where we are using base 2 logs and $\theta_d^{(i)}$ represents the ith component of vector $\boldsymbol{\theta}_d$. Using base-2 logs, the divergence between two documents is between 0 (for $\boldsymbol{\theta}_d = \boldsymbol{\theta}_{d'}$) and 1 (for $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_{d'}$ with disjoint non-zero components).

Regarding the computation of the semantic graph, evaluating (1) for all pairs of documents can easily become unfeasible for very large datasets. A first possibility to facilitate the calculation is to use parallel computing. For further scalability we use a bound on the Jensen-Shannon based on Hellinger distance, namely

$$\mathrm{JS}(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}) \geq 1 - \sum_i \sqrt{\theta_d^{(i)} \theta_{d'}^{(i)}}. \tag{2}$$

Since (2) is more economic to evaluate than (1), we can use this bound to skip the calculation of Jensen-Shannon divergences above a predetermined threshold.

Once the semantic graph has been constructed, we still need to find papers that share a high density of connections among them. For that, we will use a well-known algorithm for community detection in graphs, the Louvain algorithm (Blondel et al., 2008), that pursues to maximize an objective criterion known as modularity, which measures

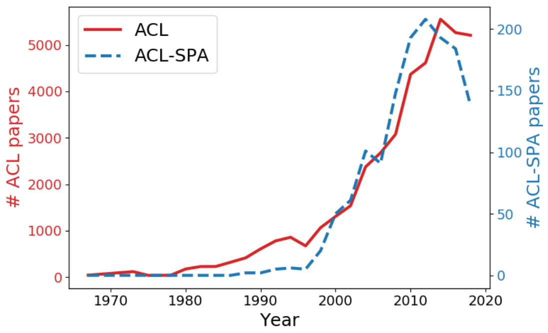

Figure 1: Papers per 2-year interval (production is larger in odd years given frequency of certain events). Spanish contribution is also represented in dashed blue (note the different scaling of the y-axis)

precisely the density of links inside identified communities with respect to the density of links between nodes across different communities. In order to visualize the graphs and apply the Louvain algorithm, we have used Gephi.

3 *Experiments*

In this section we present the results of the analysis, and discuss the thematic composition and evolution of the HLT field according to the ACL Anthology papers. We depict also the role of the contribution from Spanish researchers.

The section starts with a description based on metadata features. We continue our discussion presenting the results from the topic model, and analyzing the relative importance of ACL-SPA papers in the identified topics. The last subsection presents the results from the graph analysis, describing the detected communities, and their time evolution using graph representations.

3.1 Overall perspective of the Spanish contribution

We start analyzing the volume of the scientific contributions in the field over time. Fig. 1 shows that scientific production has increased very significantly from year 2000, showing approximate exponential increase. Spanish contribution shows approximately the same trend, with presence of Spanish authors in approximately 3.5% of the papers used for the study.

Next, we analyze the scientific production of Spanish institutions. For this, we consider a paper is assigned to an institution if any of the authors is declaring such affiliation. Affiliation detection is based on the email addresses provided by the authors in the paper. We also applied some rules for equivalent domains (e.g., `ub.edu` is equivalent to `ub.cat`). Fig. 2 shows the number of contributions of the 10 most active Spanish institutions in the ACL Anthology. These institutions alone are responsible for approximately 75% of available contributions in the ACL-SPA subset. Most relevant private companies are Barcelona Media, Vicomtech and Telefonica I+D with 20, 17, and 16 identified papers, respectively.

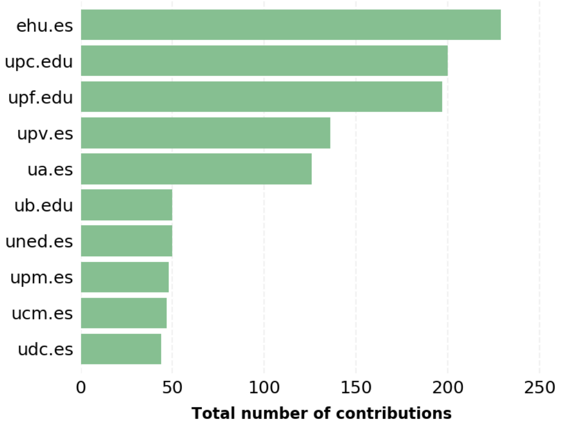

Figure 2: Number of contributions from most active Spanish institutions

3.2 Topic distribution of contributions

We have created an LDA topic model using $K = 25$, as an appropriate tradeoff between topic granularity and number of meaningless topics. All topics have been checked and annotated considering both the most significant words for each topic and the output of an automatic annotation algorithm. One topic was removed from the model because of a very poor representation in terms of most significant words (mostly typos). Table 1 describes the 24 restating topics. We can see that, in most cases, LDA found very relevant topics for the field, providing a good insight into the different subfields in HLT.

Figure 3 shows also topic size illustrating the relative contribution of Spanish institutions. Spanish sector has a relative larger size for topics 3, 4, 9, 12, and 16. Observing topic description in Table 1, many of these topics have to do with the creation of

tpc	Size	Description	tpc	Size	Description
0	0.075	supervised_learning, deep_learning,...	12	0.039	machine_translation,...
1	0.058	classification, supervised_learning,...	13	0.039	embedding, deep_learning,...
2	0.057	computational_linguistics,...	14	0.037	part-of-speech, multiword...
3	0.053	resource, annotation, corpus,...	15	0.037	natural_language_generation,...
4	0.050	resource, format, platform	16	0.035	lexical_resources, lexico-...
5	0.048	question_answering, summarization,...	17	0.035	sentiment_analysis, opinion_mining,...
6	0.045	computational_semantics, reasoning,...	18	0.035	speech_processing, dialog_system,...
7	0.044	grammar, parsing, treebank	19	0.032	information_extraction, BioNLP,...
8	0.042	parsing, dependency_grammar, treebank	20	0.030	speech_recognition, transcription,...
9	0.041	semantics, word_sense_disambiguation	21	0.029	parallel_corpora, alignment,...
10	0.040	tagger, part-of-speech,...	22	0.026	morphological_analysis,...
11	0.039	information_retrieval,...	23	0.025	computational_pragmatics, discourse,...

Table 1: Description of topic model with 24 topics

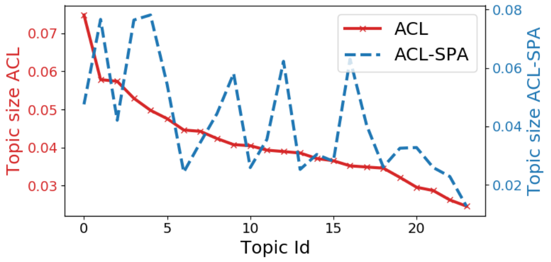

Figure 3: LDA topic size for the ACL corpus and estimated topic size for ACL-SPA papers

HLT resources (most likely for co-official languages in Spain). Topic 12 regarding Machine translation shows also a representation clearly above the average size of the Spanish sector as a whole.

3.3 Evolution of graph communities

Different experiments were conducted to select the best results for community detection and graph analysis. Community detection and graph analysis revealed to have better results when using the 40 topic-model based on the abstracts with a modularity of 1 (distinct values could be used for changing community size or radius) compared to those based on topic-models using the full paper.

Analyzing the data of the detected communities and the generated graph, 21 relevant communities were clearly identified reflecting the main areas of HLT. Each included an average of 1600 paper-entries. These communities were validated manually to assign a Community Description for each (see Table 2). It is interesting to point out that Evaluation, Validation and Quality Measures were clearly detected in a community. Parallel bilingual and multilingual resources were also detected within one community together with alignment, translation memories, etc. It is also interesting to observe a detected community of NLP toolkits and platforms or a community with clear presence of the Biomedical domain.

Figure 4 shows the evolution of some of the main areas of HLT such as Grammars, Statistical Machine Translation, Language Resources, Sentiment Analysis.

Semantics, Word Sense Disambiguation, Named-Entities are clearly represented in a highly populated community with over 3800 paper-entries. Evolution of this subgraph shows how Word Sense Disambiguation and Word2Vec started to populate more areas in this community in the last years (See Figure 5). The Spanish participation is represented in blue nodes, where it is clear that Spanish groups are more active in Semantics, WordNet and Ontologies. Moreover, the graph analysis allows zooming up into certain communities. Figure 6 illustrates a zoom of the communities related to Statistical Machine Translation. Evaluation of MT and Proceedings of the Workshop on MT (WMT) are detected as sub communities.

4 Conclusions

In this work, we have landscaped the HLT Sector through scientific literature included in the ACL Anthology, considered as one of the most important resources for HLT.

The methodology for the analysis heavily relies on Topic Modeling and Graph Analysis tools. Our approach allows to automatically detect the most significant thematics, as well as learning the semantic relation among pa-

#Papers	Community Description
3,858	Semantics, WSD, Anaphor, Semantic_Role_Labeling, Metaphor, Coreference, Syntax
3,109	Information_Extraction, Named_Entities, Temporal_Expressions, Factoids, Sentiment_Analysis
3,027	Lexicology, Semantics, Terminology, Dictionaries, WSD, Word_Vectors, Word_Embeddings
2,813	Language_Understanding, Language_Generation, Machine_Translation, Semantics, WSD, MWEs
2,074	Grammars, HSPG, Finite_State_Transducer, Automata, Formalism, Syntax
2,068	Parsing
2,055	Machine_Translation
1,833	Language_Modelling, Statistical_Language_Processing, HMM, Finite_State, Decision_Trees, SVM, CRF, Unsupervised_Learning, Deep_Learning, Neural_Networks, Chinese, Japanese
1,759	Language_Resources, Corpora, Multilingual_Resources, Minority_Languages
1,739	Speech_Recognition, Phonology, Acoustics, Spoken_Language, Dialog
1,541	Question_Answering, Information_Extraction, Information_Retrieval, Clustering, Cross_Language
1,526	Dialog, Speech_Processing, Interaction, NL_Understanding, Assitants, Agents
1,486	Domain_Specific, News, Email, Author_Profiling, Financial_Text, Political_Text, Narrative, Sports, Social_Media, Chat, Stance, Emotions, Opinions, Recommendation_System
1,419	Parallel_Corpora, Bilingual_Resources, Alignment, Translation_Memories, Multilingual_Resources
1,310	Discourse, Speech, Dialog, Pragmatics, Natural_Language_Generation, Anaphor
1,271	Evaluation, Validation, Evaluation_Metrics, Quality_Measures, Error_Detection
1,264	Ontology, Terminology, Information_Extraction, Taxonomy, Medical Domain, BioNLP
1,149	Summarisation, Language_Generation, Information_Extraction, Indexing, Document_Clustering
1,037	Annotation_Tools, Toolkits, Architecture, Interfaces, Text_to_Speech, Interaction
737	Morphology, Morphological_Analysis, Morphotactics, Inflection, Agglutinative, Stemming
728	Sentiment_Analysis, Subjectivity, Opinion_Mining, Polarity, Stance, Social_Media, Twitter

Table 2: Detected communities.

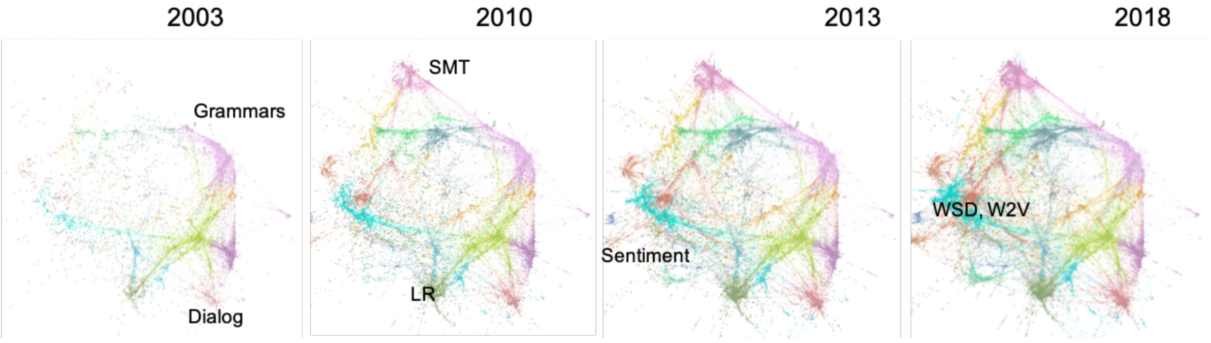

Figure 4: Evolution of main detected HLT communities (1983-2018)

pers. Semantic graphs are calculated, and communities of semantically similar papers have been detected and characterized.

Spanish contribution in the field represents 3.32% of the international scientific production represented by the ACL Anthology. According to our analysis, this participation is more significant in the areas of Lexicology & Semantics, Machine Translation, Parsing, and Speech Technologies.

These observations give insights for strategic plans to promote HLT taking into consideration the rapidly growing interest in the last years for its role in the recent advances in cognitive technologies and Artificial Intelligence. Moreover, Spanish HLT key players could contribute with a significant role in the Hispanic Language Industry.

In this study, we limited the experiments to the scientific publications in the ACL Anthology. Further studies will include more resources, especially the SEPLN journal. Also, more experiments will be conducted to landscape not only the scientific literature, but also funded R&D projects.

Acknowledgments

This work has been carried out in the framework of the Spanish State Plan for Natural Language Technologies. The work of J. Arenas-García has also been partly funded by MINECO projects TEC2014-52289-R and TEC2017-83838-R.

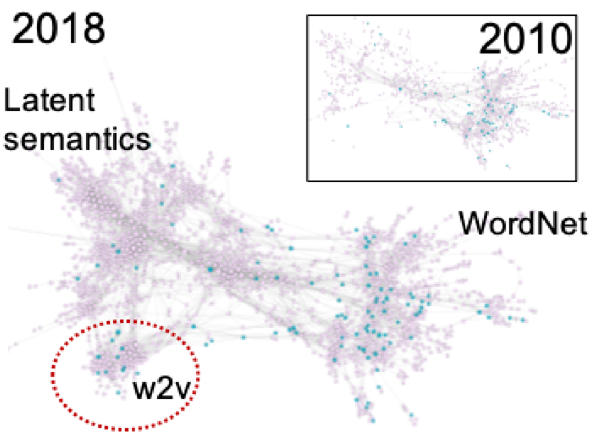

Figure 5: Latent semantic, wordEmbeddings and WordNet related papers. Contributed papers with Spanish affiliations are represented in blue

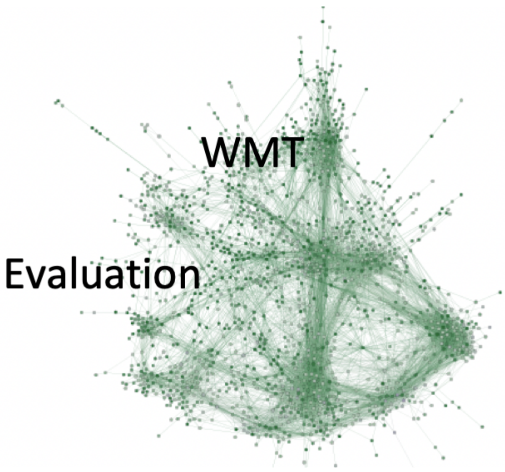

Figure 6: Graph Representation of SMT knowledge area

References

Badenes-Olmedo, C., J. L. Redondo-Garcia, and O. Corcho. 2017. Distributing text mining tasks with librAIry. In *Proc. 2017 ACM Symposium on Document Engineering*, DocEng '17, pages 63–66. ACM.

Bastian, M., S. Heymann, and M. Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks.

Bird, S., R. Dale, B. J. Dorr, B. R. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proc. LREC.*

Blei, D. M., A. Y. N, and M. I. Jordan. 2003. Latent dirichlet allocation. In *Proc. NIPS.*

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10).

Clauset, A., M. Newman, and C. Moore. 2004. Finding community structure in very large networks. *Physical review. E, Statistical, nonlinear, and soft matter physics.*

Gábor, K., H. Zargayouna, D. Buscaldi, I. Tellier, and T. Charnois. 2016. Semantic annotation of the ACL anthology corpus for the automatic analysis of scientific literature. In *Proc. LREC.*

Hall, D. L. W., D. Jurafsky, and C. D. Manning. 2008. Studying the history of ideas using topic models. In *EMNLP.*

Jin, Y., M.-Y. Kan, J.-P. Ng, and X. He. 2013. Mining scientific terms and their definitions: A study of the ACL anthology. In *Proc. 2013 Conf. Empirical Methods in Natural Language Processing.*

Langhe, R. D. 2016. Towards the discovery of scientific revolutions in scientometric data. *Scientometrics*, 110:505–519.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit.

Schäfer, U., B. Kiefer, C. Spurk, J. Steffen, and R. Wang. 2011. The ACL Anthology Searchbench. In *Proc. ACL.*

Schäfer, U., C. Spurk, and J. Steffen. 2012. A fully coreference-annotated corpus of scholarly papers from the ACL anthology. In *Proc. COLING.*

Serenko, A., N. Bontis, L. D. Booker, K. W. Sadeddin, and T. Hardie. 2010. A scientometric analysis of knowledge management and intellectual capital academic literature (1994-2008). *J. Knowledge Management*, 14:3–23.

Vogel, A. and D. Jurafsky. 2012. He said, she said: Gender in the ACL anthology. In *Proc. ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries.*

Xu, H., Z.-H. Yue, C. C. Wang, K. Dong, H. Pang, and Z. Han. 2017. Multi-source data fusion study in scientometrics. *Scientometrics*, 111:773–792.