NLP4NLP: Applying NLP to scientific corpora about written and spoken language processing

Gil Francopoulo[1], Joseph Mariani[2] and Patrick Paroubek[3]

[1] *gil.francopoulo@wanadoo.fr*
IMMI-CNRS + TAGMATICA rue John von Neumann 91405 Orsay Cedex, France

[2] *joseph.mariani@limsi.fr*
IMMI-CNRS + LIMSI-CNRS rue John von Neumann 91405 Orsay Cedex, France

[3] *pap@limsi.fr*
LIMSI-CNRS rue John von Neumann 91405 Orsay Cedex, France

Abstract
Analyzing the evolutions of the trends of a scientific domain in order to provide insights on its states and to establish reliable hypotheses about its future is the problem we address here. We have approached the problem by processing both the metadata and the text contents of the domain publications. Ideally, one would like to be able to automatically synthesize all the information present in the documents and their metadata. As members of the NLP community, we have applied the tools developed by our community to publications from our own domain, in what could be termed a "recursive" approach. In a first step, we have assembled a corpus of papers from NLP conferences and journals for both text and speech, covering documents produced from the 60's up to 2015. Then , we have mined our scientific publication database to draw a picture of our field from quantitative and qualitative results according to a wide range of perspectives: ranging from sub-domains, specific communities, chronology, terminology, conceptual evolution, re-use and plagiarism, trend prediction, novelty detection and many more. We provide here an account of the corpus collection and of its processing with NLP technology, indicating for each aspect which technology was used. We conclude on the benefits brought by such corpus to the actors of the domain and on the conditions to generalize this approach to other scientific domains.

Conference Topics
Methods and techniques, Citation and co-citation analysis, Scientific fraud and dishonesty, Natural Language Processing

1 Introduction
The NLP4NLP corpus, object of this paper, covers both the written and speech sub-domains of NLP and also encompasses a small sub-corpus in which Information Retrieval and NLP activities intersect. The corpus was made at LIMSI-CNRS (France) and contains to this day 57,235 documents from various conferences and journals with different access policies (from public to restricted). Our approach was to apply NLP tools on articles about NLP itself. We chose NLP as our first application domain because we wanted to take advantage of the fact that we are knowledgeable about the domain ourselves, and thus we would be better set to appreciate the amount of in-domain knowledge required to determine the pertinence of the results returned by automatic analysis, in particular for what concerns author names, institutions labels and acronyms, the domain terminology or the scientific concepts mentioned.

2 Existing Corpora
Among all the NLP corpora available on Internet, the ACL Anthology[1] is one of the most known because of its wide coverage in terms of time span and number of papers (more than

1 http://aclweb.org/anthology

20,000 ACL related papers[2]) and also because it provides a full access to both the metadata and the contents of the papers. Most of the papers from the site are in English and come from ACL events or journals, with a few additions from other sources like the 4,550 papers from LREC conference series[3] or the 976 articles in French or English from the TALN conference series[4]. Other sites exist like SAFFRON[5] which display results obtained by processing the content of the ACL Anthology, the LREC or CLE conference sites, or the site from University of Michigan by the CLAIR group[6] is more focused on ACL and provides search functionalities supported by apparently more elaborate numerical computations. If these sites are very valuable resources for the community, they offer publications mainly focused on the processing of written material. Since the conferences on speech processing (the other "side" of the NLP domain) are mostly managed by two large associations which are ISCA[7] (for the conferences Interspeech, ICSLP and Eurospeech) and the IEEE Signal Processing Society[8] for the ICASSP conferences. To our knowledge, the previous sites mentioned constitute the main repositories of scientific articles for the NLP domain.
The respective share of papers in our corpus coming from the ACL Anthology is about only 35% of our corpus, while the remaining part is made of publications with their origin essentially from ISCA and IEEE Signal Processing. Note that although we could have limited this study to one of the two spheres, either text processing of speech, it was important for us to cover both since our lab has teams working in the two spheres and we are particularly interested in comparing their evolutions and studying the links between the two sub-domains of NLP.

3 Related works

To the best of our knowledge this is the first time that such an extensive study covering both text and speech processing domain is undertaken. From the different works that were done in the past on scientific publications, the most notable one is probably the 2012 workshop organized by ACL in Jeju (South Korea): "Rediscovering 50 years of Discoveries in Natural Language Processing". On a smaller scale and including articles in bot English and French, there is the work of Florian Boudin (Boudin 2013) on the TALN conference series. For what concerns only speech processing, there are the two recent studies presented at the occasion of the 25 years of ISCA during Interspeech 2013 (Mariani et al 2013) and more specially on resources and evaluation for text and speech processing ther is the study presented for the 15 years of LREC at LREC-2014 (Mariani et al 2014). To the best of our knowledge this is the first time that such an extensive study is undertaken.

4 Data collection

In our study we distinguish the notion of sub-corpora specific to a journal or a conference series (for instance COLING), which can also be divided according to time, using a year as unit. The combination of both filtering criteria (sub-corpus and year) identifies what we call an "event". For each document, we process to kind of information, the metadata and the textual content. Often different version of the metadata were available, which enabled to perform consistency checks. In our database, the metadata is made of the corpus name, the

2 The figures given here were valid on March 2015.
3 http://www.lrec-conf.org
4 http://www.atala.org/-Conference-TALN-RECITAL
5 http://saffron.insight-centre.org
6 http://clair.eecs.umich.edu/aan/index.php
7 http://www.isca-speech.org/iscaweb
8 http://www.signalprocessingsociety.org

year, the authors (with the given name(s) well identified from the family name(s)), and the document title.
Metadata have been cleaned by an automatic processing and manually checked by experts of the domain, limiting the checks to only the most frequent phenomena for the cases when the task was too daunting. The metadata can be considered "cleaner" as the ones generally available, in a sense that we fixed in general various typos and inconsistencies from the version publicly available. As an extra resource, we also have the ISCA member registry for speech processing papers. This registry is very useful for authors gender statistics as it contains explicit information whether the author is male or female, thus giving us the means to disambiguate epicene given names. Note that in the case of LREC, a manual identification of gender has been done for authors with an epicene given name.
Originally, the textual content of the publications is in PDF format, of two kinds: first PDF holding only a scan of the original document, without any direct access to the content in raw test format, second PDF from which the text of the original document is retrievable directly.
For the former we had to use OCR to recover the text content, see the *preprocessing* section below.

5 Data collection

Up to now we have collected 32 sub-corpora in our NLP4NLP corpus. Their list is given in table 1. In the corpus, the vast majority (90%) of the documents comes from conferences and the remaining part from journals. As a convention, we call "document", an article which has been published in a given conference or journal and we call "paper", the physical object which holds a unique identifier. The difference is subtle, as we will see. In fact, it could be observed that the total of the cells of the table does not give exactly a grand total of 57,235 documents but slightly more (59,766) because a small number of conferences are joint conferences for some years, which means that a single paper matches with two different documents which respectively belong to two different corpora. Quantitatively, this is not an important phenomenon, because joint events happen relatively rarely, but these situations makes comparing two sub-corpora more complex. Initially, texts are in four languages:
English, French, German and Russian. The number of texts in German and Russian is less than 0.5%. They are detected by the automatic language detector of the industrial pipeline TagParser[9] and discarded. The texts in French are a little bit more numerous (3%, 1871 exactly). They are kept with the same status as the English ones. This is not a problem because our NLP pipeline we use is bilingual.

6 Preprocessing and normalization

Textual contents and metadata are built independently in parallel. For PDF documents, we use PDFBox[10] in order to extract the text content from the articles. When the PDF document holds only a scan of the original document, we apply OCR through the Tesseract[11] application. The texts resulting from both types of conversion are encoded in Unicode-UTF8. A filtering program is applied to process the most frequent OCR problems identified. An end-of-line processing is run with TagParser dictionary in order to distinguish caesura and composition hyphenation. Then, a set of "pattern matching" rules are applied to separate the abstract, the body and the reference section. For the metadata, the author name and the title are extracted from the conference program or the BibTeX material, depending on the source. Each author name is split into a given name and a family name with an automatic check against a large given name ISO-LMF (ISO-24613) dictionary comprising 74,000 entries.

9 www.tagmatica.com
10 https://pdfbox.apache.org
11 https://code.google.com/p/tesseract-ocr

Table 1. Table List of subcorpora contained in the NLP4NLP corpus.

short name	# docs	format	long name	Language	access to content	Period	# venues [12]
acl	4262	conference	Association for Computational Linguistics conference	English	open access*	1979-2014	36
alta	262	conference	Australasian Language Technology Association	English	open access*	2003-2014	12
anlp	329	conference	Applied Natural Language Processing	English	open access*	1983-2000	6
cath	932	journal	Computers and the Humanities	English	private access	1966-2004	39
cl	777	journal	American Journal of Computational Linguistics	English	open access*	1980-2014	35
coling	3833	conference	Conference on Computational Linguistics	English	open access*	1965-2014	21
conll	789	conference	Computational Natural Language Learning	English	open access*	1997-2014	17
csal	718	journal	Computer Speech and Language	English	private access	1986-2015	29
eacl	900	conference	European Chapter of the ACL	English	open access*	1983-2014	14
emnlp	1708	conference	Empirical methods in natural language processing	English	open access*	1996-2014	19
hlt	2080	conference	Human Language Technology	English	open access*	1986-2013	18
icassps	9023	conference	IEEE International Conference on Acoustics, Speech and Signal Processing - Speech Track	English	private access	1990-2014	25
ijcnlp	899	conference	International Joint Conference on NLP	English	open access*	2005-2013	5
inlg	199	conference	International Conference on Natural Language Generation	English	open access*	1996-2012	6
isca	17592	conference	International Speech Communication Association conferences (Eurospeech, ICSLP, Interspeech)	English	open access	1987-2014	27
jep	507	conference	Journées d'Etudes sur la Parole	French	open access*	2002-2014	5
lre	276	journal	Language Resources and Evaluation	English	private access	2005-2014	10
lrec	4550	conference	Language Resources and Evaluation Conference	English	open access*	1998-2014	9
ltc	299	conference	Language and Technology Conference	English	private access	2009-2013	3
modulad	232	journal	Le Monde des Utilisateurs de L'Analyse des Données	French	open access	1988-2010	23
muc	149	conference	Message Understanding Conference	English	open access*	1991-1998	5
naacl	1000	conference	North American Chapter of ACL	English	open access*	2001-2001	10
paclic	1040	conference	Pacific Asia Conference on Language, Information and Computation	English	open access*	1995-2014	19
ranlp	363	conference	Recent Advances in Natural Language Processing	English	open access*	2009-2013	3
sem	752	conference	Lexical and Computational Semantics / Semantic Evaluation	English	open access*	2001-2014	7
speechc	549	journal	Speech Communication	English	private access	1982-2015	34
tacl	92	journal	Transactions of the Association of Computational Linguistics	English	open access*	2013-2015	3
tal	156	journal	Revue Traitement Automatique du Langage	French	open access	2006-2013	8
taln	976	conference	Traitement Automatique du Langage Naturel	French	open access*	1997-2014	18
taslp	2659	journal	IEEE Transactions on Audio, Speech and Language Processing	English	content not yet included	1993-2015	23
tipster	105	conference	Tipster DARPA text program	English	open access*	1993-1998	3
trec	1756	conference	Text Retrieval Conference	English	open access	1992-2014	23
Total	59766					1965-2015	515

Then a matching process is applied between different metadata records in order to normalize author names textual realization (e.g. matching initial with first name or normalizing compound first name typography). The result is then manually checked by some members of the team who is familiar with the domain, limiting this manual check to the most frequent items if the number of items to validate becomes too large. Then, comes an important step: the calibration of the parsing pipeline. For each corpus, an automatic parsing is performed with TagParser for identifying the presence of unknown words in the documents. We make the hypothesis that the number of unknown words, according to the number of words of the texts is a good reverse indicator of the average quality of the initial data and of the processing the material has been submitted to so far. Discrepancies in the statistical profile is used to identify subcorpora which differ too widely form the average profile. We assume that the lower the percentage of unknown words is, the better the quality of the produced text is. The calibration permits also to make modifications in the preprocessing steps and to compare quantitatively the various processing steps to ensure homogeneity of the data produced. We tried different

12 This is the number of venues where data was obtainable. There may have been more venues.

tools, like ParsCit[13] or hand-written rules, and the calibration showed that computing names, titles and content globally and directly from the PDF is a bad choice with regards to the resulting quality. This is why we do not build anymore the metadata from the PDF file but from other sources.

7 Computing analysis indicators

The various analysis indicator that we produce are the following.

Basic counting: it is number of authors, one of the most basic indicators to follow the chronological evolution of each subcorpus. The number of different authors is 43,365 for 515 events.

Co-authoring counting: the aim is to follow the number of co-authors along the time line. The results show that this number is constantly increasing regardless of the corpus. Over the whole archive, the average number of co-authors varies from 1.5 for the Computer and the Humanities journal, to 3.6 for LREC. Some additional counting are made concerning the signature order: is an author's always or never mentioned as first author?

Renewal rate: This indicator shows the author turnover. It asnwers the question whether the community associated to a subcorpus is stable or not.

Gender counting: the author sexual gender is determined from the given name together with a member registry for ISCA and LREC for authors with an epicene given name. The goal is the study the proportion of men and women with respect to time and subcorpus.

Geographical origin: for a certain number of corpora, we have access to affiliations and we are able to compute and compare the distribution of the organizations, countries and continents.

Collaboration studies: a collaboration graph is built in order to determine the cliques and connected components in order to understand the set of authors is structured, i.e. who work with who (co-sign an article)? For each author, various scores are computed like harmonic centrality, betweenness centrality and degree centrality. We determine whether an author collaborates a lot or not, and whether an author sometimes signs alone or always signs with other authors. We compute a series of global graph scores like diameter, density, max degree, mean degree, average clustering coefficient and average path length in order to compare the structure of the communities around the different conferences and to understand whether and how the authors collaborate.

Citations: the reference sections of the documents are automatically indexed and the citation links are studied within the perimeter of the 32 corpora. The H-Index are computed for each author and conference. The differences are important, starting at 5 for JEP and 11 for TALN (French conferences) to 71 for ACL, and this point highlights the citation problem with respect to the language of diffusion. As for the collaboration study, the citation graphs are built both for papers and authors. We are then able to determine which are the most cited documents compared to the most citing ones. It is easy to compute the publication rate with respect to the citation rate with for instance Kishore Papineni who did not published a lot but whom the document proposing the BLEU score (Bilingual Evaluation Understudy) is cited 1,225 times within our corpus. The most cited author is Hermann Ney with 3 927 citations, with a self-citation rate of 16%.

Terminological extraction: the aim is to extract the domain terms from the abstracts and bodies of the texts. Our approach is called "contrastive strategy" and contrasts a specialized corpus with a non-specialized corpus in the same line as TermoStat (Drouin 2004). Two large non-specialized, one for English, one for French, were parsed with TagParser and the results were filtered with syntactic patterns (like N of N) and finally two statistical matrices were recorded. Our NLP4NLP texts were then parsed and contrasted with this matrix according to

13 https://github.com/knmnyn/ParsCit

the same syntactic patterns. Afterwards, we proceeded in two steps: first, we extracted the terms and we studied the most frequent ones in order to manually merge a small amount of synonyms which were not in the parser's dictionary. And then, we reran the system. The extracted terms are for most of them single terms (95% for LREC). In general, there are common nouns, as opposed to rare proper names or adjectives.

Bibliographic searches: we transform the result of the parser (which natively produces a PASSAGE format[14], based on ISO-MAF (ISO-24611) with additional annotations for named entities) into RDF in order to inject these triplets into the persistent storage Apache-Jena15 and thus to allow the evaluation of SPARQL queries. It should be noted that instead of processing an indexation and query evaluation on raw data, we index the content after preprocessing. The reason is threefold: 1) we avoid low level noise like caesura problems which are fixed by the preprocessing step, 2) the query may contain morphosyntactic filters like lemmatized forms or part-of-speech marks, 3) the query may contain semantic filters based on named entities semantic categories like company, city or system names. Of course, all these filters may be freely combined with metadata.

Term evolution: with respect to the time line, the objective is to determine the terms which are popular. For LREC, it is "annotation". We determine the terms which were not popular and which became popular like "synset", "XML" and "Wikipedia". Some terms were popular and are not popular anymore like "encoding" or "SGML". We also study a group of manually selected terms and compute the usage of "trigram" compared to "ngram". Let's add that there are some fluctuating terms (depending on a specific time period) like "Neural Network", "Tagset" or "FrameNet".

Weak signals: the aim is to study the terms which have a too small number of occurrences to be statistically taken into consideration but which are considered as "friends" of terms whose evolution is interpretable statistically. The notion of friend is defined by the joint presence of the term within the same abstract. Thus, we find that "synset" has friends like "disambiguation" or "Princeton".

Innovative feature: based on the most popular terms during the last years, the aim is to compute the author, the document and the conference mentioning this term for the first time. Thus, for instance, "SVM" appears in the LREC corpus for the first time in an Alex Weibel's document published in 2000. It is then possible to detect the conferences producing the most innovative documents.

Hybrid individual scoring: the aim is to compute an hybrid scoring combining: collaboration, innovation, production and impact. The collaboration score is the harmonic centrality. The innovation score is computed from a time-based formula applied to term creation combined by the success rate of the term over the years. The production is simply the number of signed documents. The impact is the number of citations. We then compute the arithmetic mean from these four scores. The objective is not to publish an individual hit parade but to form a short list of authors who seem to be important within a given conference.

Classification: from the extracted terms, it is possible to compute the most salient terms of a document from TF-IDF[15] and to compute a classification in order to gather similar documents within the same cluster. We use an UPGMA algorithm on a specific corpus. This tool is very helpful, because when we pick an interesting document, the program suggests a cluster of documents which are semantically similar, in the same vein as Amazon proposing a similar object or YouTube proposing the next video.

Plagiarism and reuse studies: we define "plagiarism" as the recall of a text written by a group of authors X by an author Y who does not belong to group X. We define "reuse" of an

14 http://atoll.inria.fr/passage

15 We define the salient terms as the five terms with the higher TF-IDF.

author as the recall of a text by himself in a posterior publication, regardless of the co-authors. In a first implementation, we compared raw character strings but the system was rather silent. Now we make a full linguistic parsing to compare lemmas and we filter secondary punctuation marks. The objective is to compare at a higher level than case marking, hyphen variation, plural, orthographic variation ("normalise" vs "normalize"), synonymy and abbreviation. A large set of windows of 7 sliding lemmas are compared and a similarity score is computed. We consider plagiarism and reuse when a given level is exceeded (3% for plagiarism and 10% for reuse). Concerning the plagiarism results, we did not notice any real plagiarism (one author reusing verbatim the material of another author without any citation). In contrast, we observe sometimes some groups of authors (with an empty intersection) who apparently copy-paste large fragments of texts while engaged in common collaborations. Reuse, in contrast is more frequent.

8 Conclusion and perspectives

Up to now the NLP4NLP corpus has only been used by our team, but we plan to make the part of the corpus which has no copyright restrictions publicly available shortly. The early feedback from our legal department seems to indicate that there should be no problem for the majority of the texts (80%) because the texts and the metadata are already publicly available. In contrast, a certain number of metadata and textual contents belong to Springer or associations like IEEE, these of course we will not be able to distribute. We plan to use RDF as distribution format, so as to respect W3C recommendations concerning Open Linked Data and to be compatible with the current regional project called "Centre for Data Science"[16] (CDS). The preliminary results that the NLP4NLP corpus enabled us to extract from the indicators computed with core NLP technology (up to the level of full automatic parsing) provided quantitative assessments of facts that we knew from our knowledge of the field, e.g. the rise and fall of some terms though time. The most import lesson to draw from this first experience is the fact that the point of view and knowledge from experts of the community under study is essential to provide information that cannot be recovered automatically (e.g. sexual gender or given names of authors) and to ensure that the statistics produced do not contain too large discrepancies with respect to the actual state of the domain.

9 References

Boudin, F 2013 TALN archives: une archive numérique francophone des articles de recherche en traitement automatique de la langue. TALN-RÉCITAL 2013, 17-21 June 2013, Les Sables d'Olonne, France.

Drouin, P 2004 Detection of Domain Specific Terminology Using Corpora Comparison, in Proceedings of LREC 2004, 26-28 May 2004, Lisbon, Portugal

Mariani, J, Paroubek, P, Francopoulo, G, and Delaborde, M 2013. Rediscovering 25 Years of Discoveries in Spoken Language Processing: a Preliminary ISCA Archive Analysis, in Proceedings of Interspeech 2013, 26-29 August 2013, Lyon, France.

Mariani, J, Paroubek, P, Francopoulo, G and Hamon, O 2014 Rediscovering 15 years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis, in Proceedings of LREC 2014, 26-31 May 2014, Reykjavik, Island.

Ackknowledgements. This work was partically supported by the project REQUEST in *"Programme d'Investissement d'Avenir, appel Cloud computing & Big Data",* convention 018062-25005.

16 www.campus-paris-saclay.fr/site/Idex-Paris-Saclay/Les-Lidex/Center-for-Data-Science-Paris-Saclay