Semantic Processing of Legal Texts (SPLeT-2014)

Workshop Programme

14:00-15:00 – Introductory session

14:00-14:15 – Introduction by Workshop Chairs

14:15-15:00 – Invited Talk
Sophia Ananiadou
Adapting text mining from biology to the legal domain: what do we need?

15:00-16:00 – Paper session

15:00 –15:20
Tommaso Agnoloni, Lorenzo Bacci, Maria Teresa Sagri
Legal keyword extraction and decision categorization: a case study on italian civil case law

15:20 –15:40
Frane Šarić, Bojana Dalbelo Bašić, Marie-Francine Moens, Jan Šnajder
Multi-label Classification of Croatian Legal Documents Using EuroVoc Thesaurus

15:40 –16:00
Frantisek Cvrček, Karel Pala, Pavel Rychlý
Behaviour of Collocations in the Language of Legal Subdomains

16:00 – 16:30 Coffee break

16:30-17:10 – Paper session

16:30–16:50
Radboud Winkels, Jochem Douw, Sara Veldhoen
State of the ART: an Argument Reconstruction Tool

16:50–17:10
Tommaso Agnoloni
Network Analysis of Italian Constitutional Case Law

17:10-18:30 – Panel session
Panel on "Designing, constructing and using Legal Linguistic Resources"
Moderator: Simonetta Montemagni (Istituto di Linguistica Computazionale "Antonio Zampolli")
Panellists: Karel Pala (Masaryk University), Giulia Venturi (Istituto di Linguistica Computazionale "Antonio Zampolli"), Cristina Bosco (Università di Torino), Alessandro Mazzei (Università di Torino), Guillaume Jacquet (European Commission Joint Research Centre, JRC), Vern Walker (Hofstra University School of Law Hempstead), Daniela Tiscornia (Istituto di Teoria e Tecniche dell'Informazione Giuridica)

Workshop Organizers

Enrico Francesconi	Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR, Florence, Italy
Simonetta Montemagni	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Wim Peters	Natural Language Processing Research Group, University of Sheffield, UK
Giulia Venturi	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Adam Wyner	Department of Computing Science, University of Aberdeen, UK

Workshop Programme Committee

Kevin Ashley	University of Pittsburgh, USA
Mohammad Al-Asswad	Cornell University, USA
Anderson Bertoldi	Universidade do Vale do Rio dos Sinos, Brazil
Danièle Bourcier	Humboldt Universität, Berlin, Germany
Thomas Bruce	LII Cornell, USA
Pompeu Casanovas	Institut de Dret i Tecnologia, UAB, Barcelona, Spain
Jack Conrad	Thomson-Reuters, USA
Michael Curtotti	Australian National University, Australia
Matthias Grabmair	University of Pittsburgh, USA
Marie-Francine Moens	Katholieke Universiteit Leuven, Belgium
Thom Neale	Sunlight Foundation, USA
Karel Pala	Masaryk University, Brno, Czech Republic
Paulo Quaresma	Universidade de Évora, Portugal
Erich Schweighofer	Universität Wien, Rechtswissenschaftliche Fakultät, Wien, Austria
Rolf Schwitter	Macquarie University, Australia
Daniela Tiscornia	Istituto di Teoria e Tecniche dell'Informazione Giuridica of CNR, Florence, Italy
Tom van Engers	Leibniz Center for Law, University of Amsterdam, Netherlands
Vern R. Walker	Hofstra University School of Law, Hofstra University, USA
Radboud Winkels	Leibniz Center for Law, University of Amsterdam, Netherlands

Table of Contents

Author Index

Preface

Since 2008, the LREC conference has provided a stimulating environment for the Workshop on "Semantic Processing of Legal Texts" (SPLeT) focusing on the topics of Language Resources (LRs) and Human Language Technologies (HLTs) in the legal domain. The workshops have been a venue where researchers from the Computational Linguistics and Artificial Intelligence and Law communities meet, exchange information, compare perspectives, and share experiences and concerns on the topic of legal knowledge extraction and management, with particular emphasis on the semantic processing of legal texts. Along with the SPLeT workshops, there have been a number of workshops and tutorials focussing on different aspects of semantic processing of legal texts at conferences of the Artificial Intelligence and Law community (e.g. JURIX, ICAIL).

To continue this momentum and to advance research, the 5th edition of SPLeT has been organised in conjunction with LREC-2014. LREC provides a forum in which to report on applications of linguistic technologies to particular domains as well as a context where individuals from academia and industry can interact to discuss problems and opportunities, find new synergies, and promote initiatives for international cooperation. Thus, the workshop at LREC is expected to bring to the attention of the broader LR/HLT community the specific technical challenges posed by the semantic processing of legal texts and also share with the community the motivations and objectives which make it of interest to researchers in legal informatics.

The last few years have seen a growing body of research and practice in the field of AI & Law which addresses a range of topics: automated legal argumentation, semantic and cross-language legal IR, document classification, legal drafting, open data in the legal domain, as well as the construction of legal ontologies and their application. In this context, it is of paramount importance to use NLP techniques and tools as well as linguistic resources supporting the process of knowledge extraction from legal texts.

New to this edition of the workshop and in line with LREC 2014 Special Highlight we organized a panel on the topic of "Legal Linguistic Resources" with the final aim of constructing a map of legal language resources, enabling their reuse (in reproducing and evaluating experiments) and extension. The resources presented and discussed in the panel range from annotated corpora to lexicons, thesauri and ontologies.

We would like to thank all the authors for submitting their research and the members of the Program Committee for their careful reviews and useful suggestions to the authors. Thanks are also due to the panellists who contributed the first panel organized around the topic of legal linguistic resources. We would also like to thank our invited speaker, Sophia Ananiadou, for her contribution. Last but not least, we would like to thank the LREC 2014 Organising Committee that made this workshop possible.

The Workshop Chairs

Enrico Francesconi
Simonetta Montemagni
Wim Peters
Giulia Venturi
Adam Wyner

Legal keyword extraction and decision categorization: a case study on italian civil case law

T. Agnoloni, L. Bacci, M.T. Sagri
Institute of Legal Information Theory and Techniques
Via de'Barucci 20, 50127 Firenze
{agnoloni,bacci,sagri}@ittig.cnr.it

Abstract

In this paper we present an approach to keyword extraction and automatic categorization of italian case law of first instance on civil matters. The approach complements classic NLP based analysis of texts with legal and domain features extraction. The study originated from an experimental activity promoted by the Ministry of Justice for automated semantic metadata attribution in case law deposit in the framework of the digitalization of civil trial in Italy.

Keywords: legal terminology extraction, case law categorization, legal features support

1. Introduction

The paper reports on an experimental activity, carried on in collaboration with the Italian Ministry of Justice, for the application of language processing techniques to the jurisprudential production of the court of first instance on civil matters of a major tribunal in Italy.

In recent years significant progress have been made towards the digitalization of italian courts workflow. This is particularly true for the civil matters area, where, the introduction of the PCT (*Processo Civile Telematico* - Civil Trial Online) after a long experimentation phase is replacing the traditional paper based workflow in every italian court as the only legally valid document exchange mean.

As a result, an increasing number of digitally native documents is being produced and deposited in the local repositories of each district. However, lot has to be done to make this wealth of information accessible both to legal operators (judges and lawyers) and to the wider public. This is particularly true for case law of first instance pronounced by local courts. While for higher courts the Italian Court of Cassation maintains a central repository where case law of legitimacy is organized and published with additional metadata manually attributed by its editorial office (*ufficio del massimario* cfr. Italgiure [1]), for first instance decisions this editorial activity is unaffordable and usually supplied by private legal journals with payment services on a selection of the cases.

In this scenario, with the perspective of the creation of a publicly accessible centralized repository collecting case law from local courts, the Ministry of Justice promoted this experimental activity in order to test methodologies and implement prototype software tools able to automatically enrich documents with semantic metadata or at least support the drafter with automatic suggestions to be validated at the time of decision deposit.

The activity described here was focused on the semantic enrichment of the judgements with meaningful tags describing their content from the point of view of:

- the facts of the case (*fatto*). The circumstances, as resumed in the judgement, under which the case took place.
- the legal profile of the case (*diritto*). The legal concepts (the circumstances as regulated in legal sources) significant for the case upon which the decision is formed.

Regarding the legal profile, the requirement of the project was to test two different approaches :

- bottom up: keywords emerging from the text as explicitly written in the document;
- top down: attribution of labels, not necessarily appearing in the text, as selected from a closed set of subjects appearing in a civil domain classification scheme.

The analysis was exclusively based upon features extracted through the processing of plain text made available in a homogeneous corpus of first instance decisions (in italian) on civil matters issued by an italian court. No additional contextual information or metadata was available to the analysis.

A different profile, the relation of documents with other legal documents, namely legislative and jurisprudential documents, was the subject of a different work package of the same project and was tackled by machine processing of texts as well, by means of reference parsers specifically designed to manage italian legal sources [2], (Bacci et al., 2013).

2. Case law text analysis

The corpus made available for the study is composed by approximately 7000 documents, namely all the available deposited judgements pronounced by an italian court of first instance on civil matters over a time span of 5 years from 2008 to 2013. The documents are available in pdf format from which plain text can be extracted.

A qualitative analysis of the provided corpus was performed with the support of legal experts, showing a high

[1] www.italgiure.giustizia.it/

[2] www.xmleges.org

variability both in length of the documents and in the number of details given for the exposition of the facts and details given for the legal motivations leading to the decision. Since no fixed formal structure and no drafting rule is followed, the decisions may differ in style depending on different judges and different cases. Moreover a relevant part of the legal content of the document is only implicitly lexicalized by means of legal references to external sources.
Similarly to other legal texts, italian case law make extensive use of a specialized lexicon and complex sentences. Therefore their full comprehension can be a hard task for laymen and one of the most challenging for the machine.

2.1. NLP stack

The text analysis was approached by setting up a linguistic stack for the italian language, based on open reusable resources (like OpenNLP [3]) in a Java environment. In particular the adaptation of the part-of-speech tagger to the italian language, not available off-the-shelf, has been performed based on a training corpus made available for reuse with open license from the project Paisà (Brunello, 2011). The Paisà corpus is composed of non copyrighted texts of common italian language collected from the web and richly annotated with state of the art performance linguistic analysis tools (not available to our project, see (Attardi et al., 2010)) and partially revised manually.
The lemmatization task in our linguistic analysis stack includes "Morph-it!", a free corpus-based morphological resource for the Italian language (E. Zanchetta, 2005) which provides, among other features, a lemmatization map for a wide generic italian lexicon.
Multi-terms have also been taken into consideration. A general rule for the Italian language, based on the part-of-speech of single terms, has been used in order to extract multi-terms from generic N-Grams:

$$(Noun|Adj|Art|Prep) * Noun(Noun|Adj|Art|Prep)*$$

This linguistic pattern captures any sequence and subsequence of consecutive nouns, adjectives, articles and prepositions with at least a noun in it. Generic keywords are selected with a classical statistical approach, namely the well known *tf-idf* measure which attributes a higher score of key-ness to terms that are more frequent in the input document and less spread over the different documents of the whole corpus, as they are best candidates to represent the document itself. The standard *tf-idf* formula has been used for single-term ranking, while a modified version of *tf-idf*, based on the work (Panunzi et al., 2008), has been implemented in order to rank multi-terms.
As a result of the statistical approach applied over the given corpus of decisions, the emerging keywords are often related to the facts of the case, specific to the single decision, more than to the legal aspects and legal motivations, a terminology more common and spread among many decisions.
Since our interest was in extracting both the terminology concerning the facts and the terminology concerning the legal aspects, and since the former was covered by the statistical approach, for the latter we opted for the exploitation of an external linguistic resource specialized on the legal domain in order to apply legal relevance weights to single and multi terms.

[3] http://opennlp.apache.org/

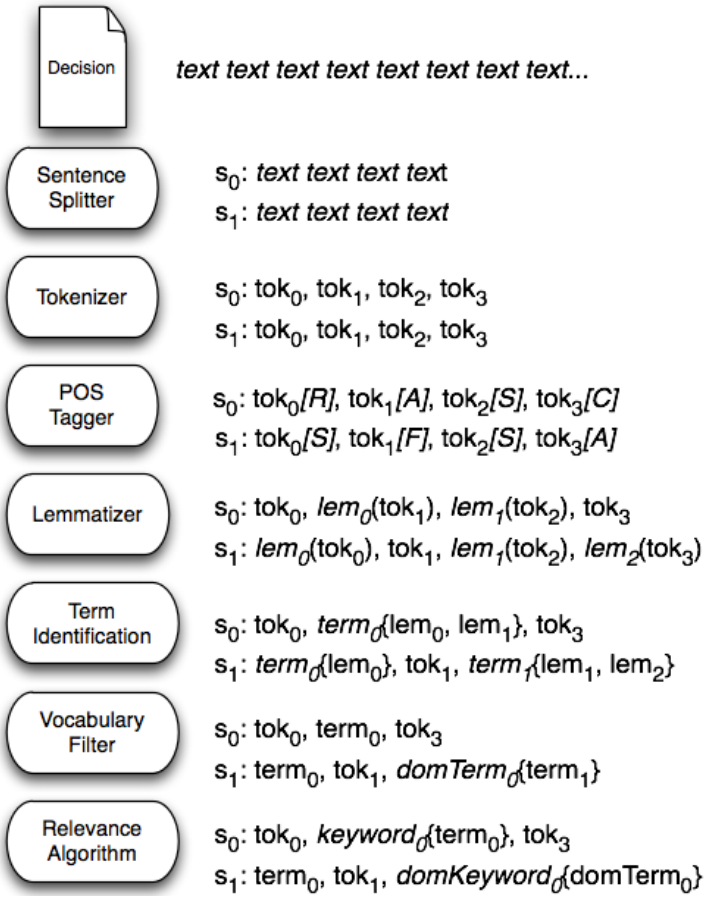


Figure 1: The NLP stack.

2.2. Vocabulary filter

Judges and legal scholars are typically more interested in the legal aspects of a decision, and they are used to categorize cases with criterions based on those legal aspects more than on the peculiar facts of a case. Therefore, in order to filter and re-rank the legal terminology of a document, our practical approach consisted in exploiting as much as possible the available terminological resources specific to the legal domain.
A Vocabulary Filter layer is hence implemented in our analysis stack. It is fed by both existing domain vocabularies and by domain terminology automatically extracted from selected relevant legal sources eventually revised by legal experts. Its purpose is the identification of the legally relevant terminology (both single and multi terms) with semantic relevance as descriptor of the legal subject of a case.
The resources included in the vocabulary filter are:

- JurWordnet, a legal lexicon developed at ITTIG (M.T. Sagri, 2003), with a wordnet-like structure;
- an automatically extracted lexicon from the civil code,

which is the systematic collection of laws dealing with the core areas of civil law matters and represents the main legal source that applies over the decisions of our corpus;

- an automatically extracted lexicon from the titles *(rubriche)* of the partitions of the civil code. This is a subset of the previously described lexicon but its terms are semantically heavier: with the aid of legal experts we evaluated that the text included in titles is practically always explicitly related to the legal domain and constitutes a rich source of legally relevant terminology on civil matters. We were able to isolate from the whole text the title of each partition by exploiting the formal structure of legal sources explicitly annotated in XML starting from the plain text (see sect. 2.3.).

Lexicon	Size
JurWordnet	7912
Civil Code Titles	6323
Civil Code	18841

Table 1: Number of terms in each resource

The overall resource has been used to identify the legal domain terminology in the text of decisions, hence producing a *bag of legal domain terms*. Such terms are then ranked depending on:

- their frequency in the document;
- the lexicon they belong to: since JurWordnet is a resource entirely built by legal expert, it is considered the most reliable, so its terms weigh more than the terms from the titles lexicons and more than the ones from the overall civil code lexicon.

The identification of the legal terminology and a ranking method for legal terms make possible to select and extract legal domain keywords from the text of a decision.

2.3. Legal features support

In legal informatics and increasingly in official publications, the formal structure of a legal text is grasped and explicitly annotated with Legal XML formats (Nir format in Italy).

Being no official XML publication of norms yet available in Italy, in order to produce a lexicon of the titles of the civil code we retrieved it from the web in plain text and then used legal features extraction tools available for Italian legal texts developed at ITTIG in previous projects.

We used xmLegesMarker (Bacci et al., 2009), a structural parser of legislative texts, to automatically annotate the civil code text in XML-Nir format. The structural mark-up identifies the formal partitions of the text (the italian civil code is composed of almost 3000 articles grouped in 50 chapters and 6 books), numbers, titles and identifiers of the individual partitions. Though this kind of annotation carries little semantics in terms of the content of the norm, it provides a fine-grained structure of the text that

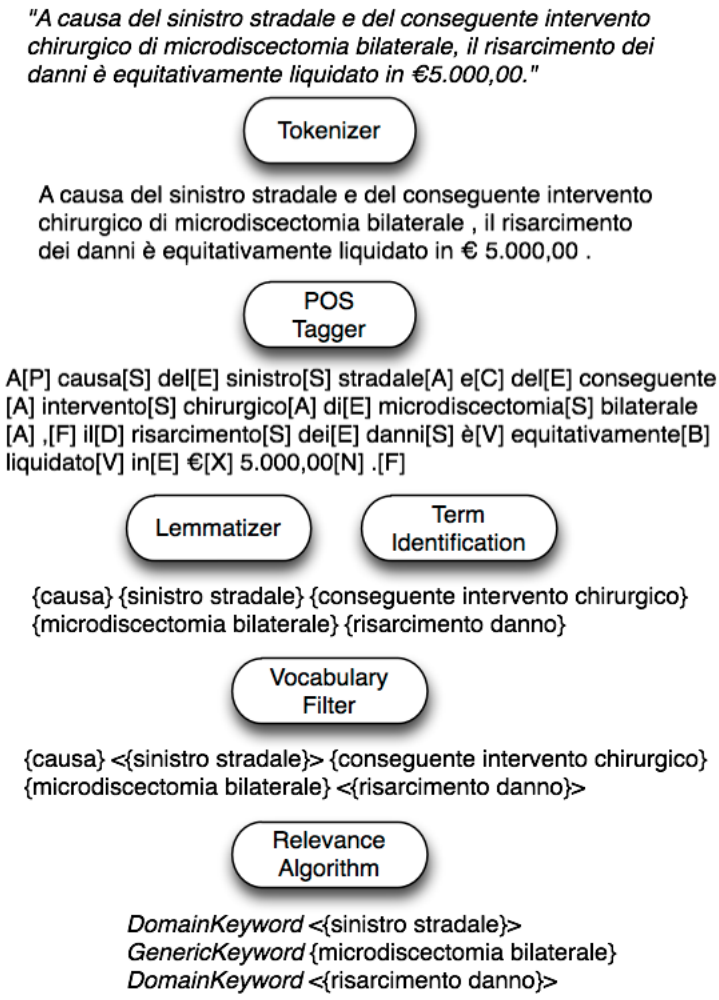


Figure 2: The NLP stack at work.

ultimately allows to apply different weight to terminology extracted from different sections (*e.g.* from the titles or from the body of partitions).

Another important legal feature used to support text analysis are legal references. Legal references appearing in the judgement are a relevant source of semantic information and sometimes (in case of poorly lexicalized texts) they carry the most part of the legally relevant semantics of the text as they are often used by judges as a technical mean to precisely refer to the legal basis underlying their decision.

xmLegesLinker is a text analysis module developed at ITTIG able to identify legislative references from legal documents in plain text. The significant fields of the citation (issuing authority, type of cited document, date, number, subpartitions) are recognized, normalized and serialized in *urn:lex* format which is the standard *de facto* for legislative sources identification.

By applying xmLegesLinker on the whole corpus the distribution of references reported in Tab. 2 emerges.

Not surprisingly from a legal point of view given the domain, legislative references to the civil code and to the civil procedure code cover more than the 80% of the total num-

Type of reference	Percentage
Civil Code	38.16%
Code of Civil Procedure	43.38%
Other references	18.46%
Total number of references	19427

Table 2: Distribution of references in the provided corpus

ber of legislative references.
The combination of references and structural markup of legal sources allows to establish a semantic link from the citing document, *e.g.* a judgement, and the knowledge and terminology carried by an external legal source. In a linked open data perspective where the legal sources are marked-up and exposed as data on the web (which is still not the case in Italy) this would allow a straightforward semantic linking among legal sources providing a rich source of knowledge as a basis for further extraction.

The introduction of these textual processing tools specifically developed for the legal domain turned out to be extremely valuable to our analysis. In particular we integrated the formal structure annotation of the civil code as a filter for domain vocabulary extraction, and the legal reference extraction as a support in the task of categorization reported in sect. 3.

3. Categorization of decisions

The experimental activity on the given corpus of first instance civil judgements included a task concerning the categorization of decisions. The categories are expected to reflect the most representative legal subjects in the given corpus with a non-trivial level of granularity. Ideally, a category will be composed of a group of decisions sharing a common legal concept. A synthetic description of the legal concept will be used as a *label* for each category. It is worth noting that labels, unlike keywords, could or could not appear explicitly in the text of a decision.
The lack of a previously annotated set of decision to be used as a training set in a classic machine-learning classification problem and the realization that the existing classification schemes would not suit the corpus well enough, brought us to consider this top-down approach.

3.1. Classification scheme definition

The most popular classification criterions for decisions on civil matters in Italy tend to focus too much either on the peculiar civil legal subject and the related social situations or on the procedural aspects of the trial. The former approach is typically used to categorize decisions in a simple and comprehensible way, but it ends up flattening out the juridical profile of the case law. The latter, more pragmatic approach, moves the focus on the aspects of the procedural *iter* of the document, but it doesn't suit the needs of judges and judicial officers well enough.
Therefore, in our top-down approach, the categories should:

- cover as much as possible the decisions in the given corpus;
- consider both legal subjects and procedural aspects of cases;
- reflect a trade-off between the granularity of the legal concepts and a number of categories sufficiently low to allow a sensible automatic assignment of decisions to categories;
- not entail a partition on the corpus: a decision can adhere to more than one category, not being the categories themselves mutually exclusive.

Given those conditions, through a legal analysis partially supported by available feature extraction tools (terminology and references) and performed by legal experts, a list of 37 categories emerged.
The perspective here was to bootstrap a process where, after the first initial effort of defining and profiling the categories, a reference scheme and a revised tagged corpus would be available to test a supervised approach for automatic classification.

3.2. Categories characterization

In order to provide a linkage criterion between the identified categories and the actual decisions in the corpus, the legal experts performed a manual work of semantic characterization of each category, that consisted in:

- a linguistic profile, a list of legal *tags* strictly pertinent to the category;

The tags characterizing each category were identified with an iterative semi-automatic process supervised by the legal domain expert: automatic domain keyword extraction from manually categorized documents, manual selection of characterizing tags from the extracted keywords and further enrichment of the profile.

- a references profile, a list of legal references known to be strongly related to the category.

In particular, the main legal sources taken into consideration for the reference profile have been the civil code and the civil procedure code, since they cover most of the overall legal references in our corpus. Thanks to xmLegesMarker and xmLegesLinker we were able to produce a normalized representation of every single partition within both the civil and the procedural code, hence being able to use partition-level references in the profile of each category.
Moreover, references to laws on specific subjects (e.g. immigration, public procurement) have also been included in the profiles of many categories, since their presence in judgements clearly evoke their belonging to that specific category.

3.3. Category assignment

Given all these premises, the task of assigning a decision to one or more category (multi-label classification, see 3.1.) reduces to providing a scoring system to rank the level of adherence of the decision against each category.
The score takes into consideration:

Category	Size	Category	Size	Category	Size
Circolazione stradale	534	Diritto di proprietà	445	Fideiussione	436
Processo di esecuzione	428	Locazione e sfratto	420	Lavoro	412
Risoluzione del contratto	403	Contratto preliminare	374	Tutela del consumatore	317
Diritto societario	307	Assicurazione	299	Abitazione e condominio	275

Table 3: Results on label assignment on the whole corpus

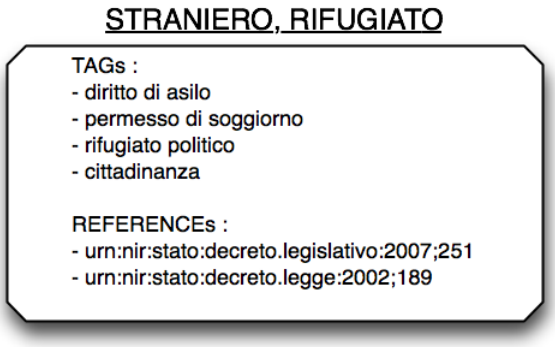


Figure 3: Extract from the profile of "Foreigner/Refugee" category

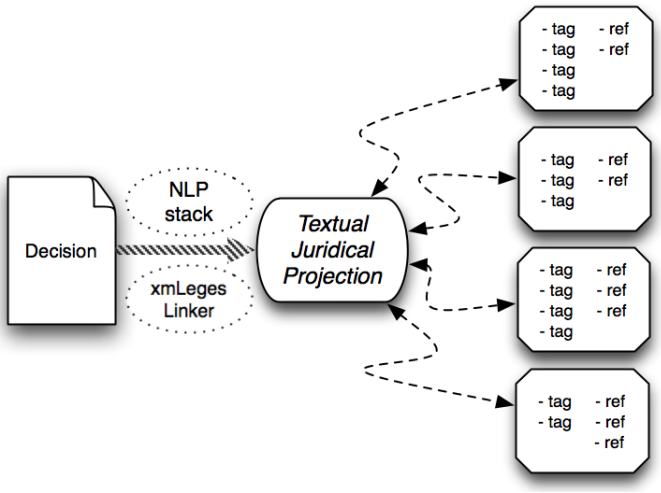


Figure 4: Scoring of the document against each category

- a measure concerning the presence of the tags in the text of the decision;
- the number of matching legal references cited by the decision.

Through the application of the NLP tools described in 2.1. and through the use of xmLegesLinker, given the text of a decision we are able to produce a *Textual Juridical Projection* of the decision, i.e. a view of the juridical components of the decision: the domain terminology and the legal references.
Such synthetic representation is then used to calculate the score:

- for every tag, partial and exact matches with terms are considered and different weights are applied depending on the length of the match, the number of words involved in the match and the belonging of a term to one of the domain lexicons described in 2.2.;
- every legal reference is represented in the *urn:lex* format, so a simple string matching is applied against every reference included in the profile of each category.

3.4. Results on categorization

The total set of documents provided by the court for the experimentation consisted of 7229 decisions on civil matters. We run the category assignment process on the whole corpus. For every decision we computed a score against every category so that:

- the labels of the categories corresponding to the three highest scores were assigned to the decision;
- a minimum score threshold was required in order to be a valid assignment.

The overall decisions that didn't make the threshold were around the 7% of the corpus. This result was expected since the provided corpus included small documents with poor legal content.
The number of valid assignments was 14504, the average number of labels on each decision 2.15, the average number of decisions assigned to each category 392. In table 3, specific numbers of assignments on a selection of categories from the classification scheme are showed.

3.5. Accuracy test

An accuracy test was run on a set composed by 328 decisions with one label assigned by a legal expert, distributed across the 37 categories of the classification scheme. Since the categorization algorithm assigns up to three labels to a decision, in table 4 we report the results when the one label assigned to a decision of the test set scores the highest, when it is either the first or the second highest score and finally when it is either the first, the second or the third highest score.

Type of test	Correct	Accuracy
Highest	223	68%
1st or 2nd highest	288	88%
1st, 2nd or 3rd highest	312	95%

Table 4: Distribution of references in the provided corpus

These results underline once more that often a decision in this kind of corpus carries more than one worthwhile legal aspect and a classification scheme with partially overlapping categories is able to capture the co-existence of more legal aspects. For 312 decisions, the pre-assigned label is found in the (three tops) labels provided by the categorization algorithm. Further validation of both the categories included in the classification scheme and of the accuracy of the documents assignment will be possible from the feedback provided by judges at the end of the experimentation phase in the court which is still ongoing.

3.6. Legal similarity of judgements

As we have extensively addressed previously, the legal aspects and motivations in texts of decisions from a first-instance court are often mixed with plenty of text about the mere facts, the actors and many other less important sentences from the point of view of the final user. The classic approach of document similarity based on comparing the vectorized version of the texts fails because of all this "textual noise". Again, we can produce the textual juridical projection of two arbitrary decisions and, with a slight modification to the previously described algorithm, calculate a score of similarity based on:

- the shared legal terminology;
- the shared legal references.

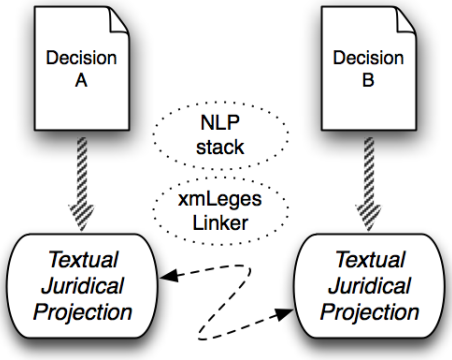


Figure 5: Scoring of the legal domain similarity between two decisions

After calculating the mutual similarity among every decision in the corpus, we managed to present, for every decision, a sorted list of the most legally similar ones, independently from the level of similarity of their non-legal aspects.

4. Conclusion

Case law is an extremely technical and specialized textual genre. The lexicon, the structure of the phrases, the implicit knowledge and overall its semantics are difficult to catch even to a human if not supported by a rich background of technical (legal) knowledge. Even more for machines. In this experiment we developed a methodology where a classical textual analysis chain, is backed up with as much as possible legal knowledge. Such knowledge is automatically extracted with legal features extraction tools developed over the years by our institute, obtained by processing selected legal resources and supervised by legal experts.
The NLP analysis chain which was a precondition to our project and not its focus, was developed internally due to the chronic lack of free and open analysis tools for the italian language (though there are some encouraging steps in this direction, if only for the open resources that we could use to train our stack).
The problem has been faced by multiple point of views. From a purely bottom-up/textual side, with the extraction of the legal and factual keywords better describing the content of texts as written by the judge in his decision. And from the higher level of abstraction of legal categories and subjects. The linkage of case law texts to the legal sources supporting the decision, which is commonly done "in the real world" by judges in their writings with the heavy recourse to textual legal references, once automated with *ad hoc* legal features extraction tools, is able to provide the machine a rich source of semantic knowledge to support its processing.
This proved to be an efficient methodology in order to meet the user information needs by automatically extracting the kind of information that users are most familiar with.
Moreover, the manual check and validation provided by judges through the integration of these supporting tools in a user interface, allows to iteratively enlarge, refine and make more reliable the domain knowledge collected in corpora over which machine learning approach could be eventually tested.

5. References

G. Attardi, S. Dei Rossi, and M. Simi. 2010. The tanl pipeline. In *Proc. of LREC Workshop on WSPP*, Malta.

L. Bacci, P. Spinosa, C. Marchetti, and R. Battistoni. 2009. Automatic mark-up of legislative documents and its application to parallel text generation. In *LOAIT 2009 - 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques*, Barcelona (E).

L. Bacci, E. Francesconi, and M.T. Sagri, 2013. *A Proposal for Introducing the ECLI Standard in the Italian Judicial Documentary System*, pages 49–58. IOS Press, Amsterdam (NL).

M. Brunello. 2011. Paisà - a creative commons corpus. In *NLP group meeting, University of Leeds*, Leeds, UK. www.corpusitaliano.it.

M. Baroni E. Zanchetta. 2005. Morph-it! a free corpus-based morphological resource for the italian language. In *Proc. of Corpus Linguistics*, University of Birmingham, Birmingham, UK. http://dev.sslmit.unibo.it/linguistics/morph-it.php.

D. Tiscornia M.T. Sagri. 2003. Metadata for content description in legal information. In *Knowledge, Ontology, Metadata And Meaning Matters, (KOM3), DEXA 2003*. Springer.

A. Panunzi, M. Fabbri, and M. Moneglia. 2008. Multilingual open domain key-word extractor proto-type. In *Proceedings of EURALEX 2008*, Barcelona, Jul. 15-19.

Multi-label Classification of Croatian Legal Documents Using EuroVoc Thesaurus

Frane Šarić*, Bojana Dalbelo Bašić*, Marie-Francine Moens†, Jan Šnajder*

*University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
frane.saric@gmail.com,{bojana.dalbelo,jan.snajder}@fer.hr
†Department of Computer Science, KU Leuven, Celestijnenlaan 200A, Heverlee 3001, Belgium
sien.moens@cs.kuleuven.be

Abstract

The automatic indexing of legal documents can improve access to legislation. EuroVoc thesaurus has been used to index documents of the European Parliament as well as national legislative. A number of studies exists that address the task of automatic EuroVoc indexing. In this paper we describe the work on EuroVoc indexing of Croatian legislative documents. We focus on the machine learning aspect of the problem. First, we describe the manually indexed Croatian legislative documents collection, which we make freely available. Secondly, we describe the multi-label classification experiments on this collection. A challenge of EuroVoc indexing is class sparsity, and we discuss some strategies to address it. Our best model achieves 79.6% precision, 60.2% recall, and 68.6% F1-score.

Keywords: multi-label classification, legislative documents, EuroVoc thesaurus

1. Introduction

Semantic document indexing refers to the assignment of meaningful phrases to a document, typically chosen from a controlled vocabulary or a thesaurus. Document indexing provides an efficient alternative to traditional keyword-based information retrieval, especially in a domain-specific setting. As manual document indexing is a very laborious and costly process, automated indexing methods have been proposed, ranging from the early work of Buchan (1983) to the more recent system by Montejo Ráez et al. (2006).

The practical value of indexing legal documents has long been recognized. Acknowledging this fact, the EU has introduced EuroVoc (Hradilova, 1995), a multilingual and multidisciplinary thesaurus covering the activities of the EU, used by the European Parliament as well as the national and regional parliaments in Europe.[1] The EuroVoc thesaurus contains 6797 indexing terms, so-called *descriptors*, arranged into 21 different fields.[2] The thesaurus is organized hierarchically into eight levels: levels 1 (fields) and 2 (microthesauri) are not used for indexing, while levels 3–8 contain the descriptors. The EuroVoc thesaurus exists in 23 languages of the EU.

In this paper we describe the work on EuroVoc indexing of Croatian legislative documents. Most of this work has been carried out within the CADIAL (Computer Aided Document Indexing for Accessing Legislation) project,[3] in collaboration with the Croatian Information-Documentation Referral Agency (HIDRA). The overall aim of the CADIAL project was to enable public access to legislation. To this end, a publicly accessible semantic search engine has been developed.[4] Furthermore, a computer-aided document indexing system eCADIS has been developed to speed up semantic document indexing. For more details about the CADIAL project, see (Tadić et al., 2009).

The focus of this paper is the machine learning aspect of the problem. Namely, EuroVoc indexing is essentially a multi-label document classification task, which can be addressed using supervised machine learning. The contribution of our work is twofold. First, we describe a new, freely available and manually indexed collection of Croatian legislative documents. Secondly, we describe EuroVoc multi-label classification experiments on this collection. A particular challenge associated with EuroVoc indexing is class sparsity, and we discuss some strategies to address it. Another challenge, as noted by Steinberger et al. (2012), is that document classification is generally more difficult for Slavic languages due to morphological variation, and we also consider ways to overcome this. Although we focus specifically on EuroVoc indexing of documents in Croatian language, we believe our results may transfer well to other languages with similar document collections.

2. Related Work

Most research in supervised learning deals with single label data. However, in many classification tasks, including document and image classification tasks, the training instances do not have a unique meaning and therefore are associated with a set of labels. In this case, multi-label classification (MLC) has to be considered. The key challenge of MLC is the exponentially-sized output space and the dependencies among labels. For a comprehensive overview, see (Zhang and Zhou, 2013; Tsoumakas and Katakis, 2007).

EuroVoc indexing can be considered a large-scale MLC problem. Mencía and Fürnkranz (2010) describe an efficient application of MLC in legal domain, where three types of perceptron-based classifiers are used for EuroVoc indexing of EUR-Lex data.

The most common approach to cope with large-scale MLC is to train a classifier for each label independently (Tsoumakas et al., 2008). Boella et al. (2012) use such an approach in combination with a Support Vector Machine

[1] http://eurovoc.europa.eu/
[2] Data is for EuroVoc version 4.31, used in this work.
[3] http://www.cadial.org/
[4] http://cadial.hidra.hr/

(SVM) for EuroVoc MLC of the legislative document collection JRC-Acquis (Steinberger et al., 2006). Steinberger et al. (2012) present JEX, a tool for EuroVoc multi-label classification that can fully automatically assign EuroVoc descriptors to legal documents for 22 EU languages (excluding Croatian). The tool can be used to speed up human classification process and improve indexing consistency. JEX uses a profile-based category ranking technique: for each descriptor, a vector-space profile is built from the training set, and subsequently the cosine similarity between the descriptor vector profile and the document vector representation is computed to select the k-best descriptors for each document. Daudaravicius (2012) studies the EuroVoc classification performance on JRC-Acquis on three languages of varying morphological complexity – English, Lithuanian, and Finish – as well as the influence of document length and collocation segmentation. Whether linguistic preprocessing techniques, such as lemmatization or POS-tagging, can improve classification performance for highly inflected languages was also investigated by Mohamed et al. (2012). Using JRC JEX tool on parallel legal text collection in four languages, they showed that classification can indeed benefit from POS tagging.

3. Croatian Legislative Document Collection

3.1. Collection Description

The document collection we work with is the result of the CADIAL project and consists of 21,375 legislative documents of the Republic of Croatia published before 2009 in the Official Gazette of the Republic of Croatia (*Narodne Novine Republike Hrvatske*). The collection includes laws, regulations, executive orders, and law amendments. The collection has been manually indexed with descriptors from EuroVoc and CroVoc. The latter is an extension of EuroVoc compiled by HIDRA, consisting of 7720 descriptors covering mostly names of local institutions and toponyms. Overall, the combined EuroVoc-CroVoc thesaurus consists of 14,547 descriptors.

The manual indexing was carried out in two rounds. In the first round, carried out before 2007, a total of 9225 documents were manually indexed. This part of the collection was used to train the machine learning-based indexer eCADIS. In the second round, carried out from 2007 onward, additional 12,510 documents were indexed (1187 international treaties, 7129 law amendments, and 4194 additional laws, regulations, and executive orders). To speed up the procedure, in this round the eCADIS indexer was used as a starting point for manual indexing. Subsequently, each document has been manually inspected and the descriptors were revised where necessary. Also, descriptors from the first round were checked and some were revised.

The law amendments have not been indexed, as they inherit the descriptors of the main regulation they refer to. We therefore did not consider law amendments in our experiments. The final collection that we use consists of 13,205 manually indexed documents, which amounts to 332K unique words and 39.9M tokens. The average document size is about 3K words. We refer to this collection as the NN13205 collection.[5]

[5]Available under CC BY-NC-SA 3.0 from `http://takelab.fer.hr/data/nn13205`

3.2. Indexing Principles and Quality

The NN13205 collection was indexed by five professional documentalists according to strict guidelines established by HIDRA. The main principle was to choose descriptors that are likely to match the end-users' information needs. This transferred to two criteria: specificity and completeness. Specificity means that the most specific descriptors pertaining to document content should be chosen. The more general descriptors were not chosen, as they can be inferred directly from the thesaurus. Completeness means that the assigned descriptors must cover all the main subjects of the document. Essentially, the indexing followed the guidelines set by (ISO, 1985), the UNIMARC guidelines, and best practices developed in HIDRA.

At first sight, the specificity criterion might seem to imply that only the leaf descriptors are assigned to the documents. However, this is not the case, as sometimes the lower levels lack the suitable descriptor. In these cases, the indexers had to back off to a more general descriptor. Consequently, if a document is best described with a number of descriptors, some of them will be more general than the others. In fact, this happens in 23.7% of documents in the NN13205 collection. Note that this effectively introduces extra semantics: although a more specific descriptor implies all the more general ones, explicit assignment of a more general descriptor indicates that the more specific descriptors are not informationally complete for the document.

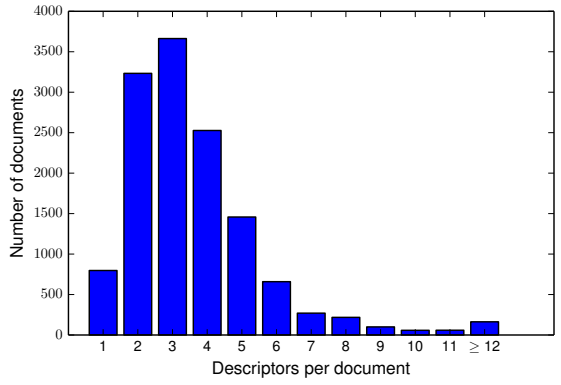
As a means of quality control, indexing has undergone periodical revisions to ensure consistency. This was done either by inspecting all documents indexed with the same descriptor or by inspecting groups of topically related documents. Unfortunately, no methodology was established to measure the inter-annotator agreement; in particular, no document was ever indexed by more than a single documentalist. As a consequence, we cannot estimate the overall quality of manual annotation using inter-annotator agreement as a proxy. Furthermore, the lack of inter-annotator estimate is troublesome from a machine learning perspective because it prevents us to establish the ceiling performance for a machine learning model on this task.

3.3. Indexing Statistics

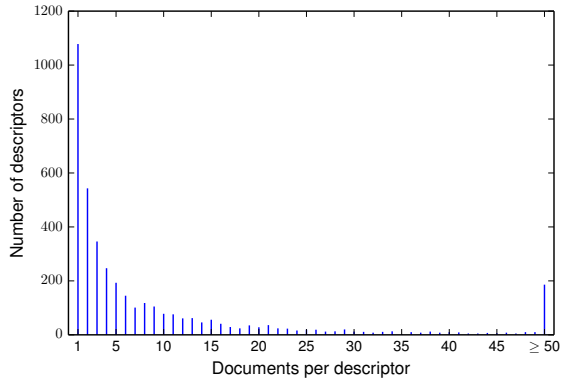
In total, 3951 different EuroVoc descriptors were used to index the 13,205 documents. Indexers typically assigned up to 10 different descriptors to each document. The total number of descriptor assignments is 48,123, which amounts to 3.6 descriptors per document (see Fig. 1a).

From a machine learning perspective, the major problem of NN13205 is that it is sparsely labeled. Out of 3951 descriptors assigned, 1078 were assigned to a single document and 2867 were assigned to less than ten documents, as shown in Fig. 1b. For comparison, the Reuters news stories corpus RCV1 (Rose et al., 2002), the benchmark collection for document classification, contains as much as 30K documents and only 100 indexing terms.

It is also interesting to compare our indexing statistics against that of the JRC-Acquis corpus (Steinberger et al., 2006). The statistics suggests that the class sparsity prob-



(a)



(b)

Figure 1: Histogram of (a) descriptors per document and (b) documents per descriptors

lem is more pronounced for the NN13205 than for the JRC-Acquis. For any single language, the JRC-Acquis has approximately 2.5 times more documents than NN13205. While NN13205 documents have anywhere from 1 to 36 assigned descriptors (avg. 3.6), JRC-Acquis documents have from 2 to 17 assigned descriptors (avg. 5.4). The total number of different descriptors in the JRC-Acquis ranges from 3584 to 4234, depending on the language.

4. Classification Experiments

4.1. Classification Model

EuroVoc indexing is essentially a hierarchical MLC problem, since each document may be assigned several descriptors. As noted in Section 2, the simplest way to address an MLC problem is to frame it as a binary classification problem, by training a separate classifier for each label. This is the approach we adopt here.

A variety of classifiers can be used for text classification. We use a Support Vector Machine (SVM) (Joachims et al., 1998), which has shown to be competitive on a wide range of classification tasks, including text classification. We use the LIBLINEAR (Fan et al., 2008) implementation, and the particular model we use is the L2-regularized L2-loss SVM. Note that we use a linear kernel, since the high number of features typical for text classification problems usually implies linear separability.

To train the binary SVM classifiers, we adopt the one-vs-rest scheme: we train a separate classifier for each EuroVoc descriptor, using documents indexed with that descriptor as the positive instances and all other documents as the negative instances.[6] If the classifier output exceeds a certain threshold, then the descriptor is assigned to the document, otherwise it is not. For improved accuracy, we additionally optimize the threshold of each individual model using the SCutFBR.1 method proposed by Yang (2001).

Another aspect that we do not explicitly consider here is hierarchy. EuroVoc indexing could be cast as a hierarchical classification problem, which has been extensively studied in the literature. A technique using combination of Bayes with SVM classifiers proposed by Cesa-Bianchi et al. (2006) shows good results, although the advantage is not so clear on real data sets. Most hierarchical models permit only leaf labels to be assigned to instances. Models have been proposed, such as the one by Sun and Lim (2001), that allow also the inner nodes to be used as labels, which is what would be required in our case because of the indexing principles used for the NN13205 collection (cf. Section 3.2.). We leave the issue of hierarchical classification for future work.

4.2. Experimental Setup

To obtain reliable error estimates and to prevent overfitting the model, we used a 5×3 nested cross-validation for model selection. Because of the large number of classifiers involved, for each model we consider only three values (1, 10, and 100) for the regularization parameter C.

We evaluate the classifiers in terms of commonly used performance measures: precision (P), recall (R), and the F1-score (the harmonic mean of P and R). Because we deal with multiple classes, we calculate the micro-average of these measures. We additionally calculate the macro-averaged F1-score (F1-score averaged over descriptors), which is more sensitive to the performance of the model on sparse classes. Note that micro P, micro R, and micro F1-score generally differ from each other because this is a multi-label problem, unlike in a multi-class (one-class-per-instance) classification problem.

As noted by Lewis et al. (2004), class sparseness raises the issue of how to compute the F1 score on under-represented classes. This has a significant impact on the overall result because NN13205 has many such classes. Stratified sampling is not an option here because the collection is sparsely multi-labeled. Instead, we decided to average the performance metrics over classes with one or more positive test instances, as proposed by Lewis et al. (2004). If, for a given descriptor, only documents from the test set are indexed with it, then a model for this descriptor cannot be trained and the F1-score is set to 0 for that descriptor. Note that this is a more realistic evaluation setting than averaging over classes with one or more positive training examples.

It should be noted that other evaluation schemes are applicable in our setting, such as the multi-label classification evaluation (e.g., Tsoumakas and Katakis (2007)) and hierarchical classification evaluation (e.g., category-similarity

[6] Subsampling negative instances, typically used to balance the classes, did not improve the performance.

Table 1: Performance on the complete NN13205 collection

Features	Micro P	Micro R	Micro F1	Macro F1
Words	**82.6**	56.5	67.1	45.9
Lemmas	80.7	**58.8**	**68.0**	47.8
Stems	80.2	58.7	67.8	**47.9**

measures proposed by Sun and Lim (2001)). We leave this line of research for future work.

4.3. Preprocessing

Prior to constructing the feature vector, we remove from each document the stop words using a manually compiled list of 2000 inflected stop words (conjunctions, prepositions, pronouns, numbers, etc.). The large number of features often poses an efficiency problem in text classification. This also applies to EuroVoc classification, where a large number of models has to be trained. To make training more efficient, we decided to employ a feature selection procedure (Yang and Pedersen, 1997). Preliminary experiments have indicated that we can discard 10% of features using the χ^2 measure without any noticeable performance loss. This leaves us with about 280K features.

Another salient problem in text classification is morphological variation, due to which a single term gets dispersed into several morphological variants. This is especially problematic for inflectionally rich Slavic languages, such as Croatian. The problem can be alleviated by morphological normalization, which for Croatian language has been shown as a useful technique for both dimensionality reduction and performance improvement (Malenica et al., 2008). In this work we experiment with two normalization techniques – lemmatization and stemming – which we apply prior to feature selection. For lemmatization, we use an automatically acquired inflectional lexicon of Croatian compiled by Šnajder et al. (2008). For stemming, we use the rule-based inflectional stemmer developed by Ljubešić et al. (2007). Lemmatization is a more accurate technique than stemming, which also takes into account the homographs by normalizing them to several lemmas. Morphological normalization reduces the number of features to ∼190K with lemmatization and ∼170K with stemming, which amounts to a reduction of about 29% and 37%, respectively.

4.4. Baseline Results

We first evaluate a model trained on the complete NN13205 collection, utilizing 3405 classifiers, one for each descriptor used. The results are summarized in Table 1. Expectedly, macro F1-score is lower than micro F1-score because the performance on sparse categories is generally lower. For the same reason, the recall is substantially lower than precision because the model generally fails to assign the rarely used descriptors. Morphological normalization improves the overall performance (4% relative performance improvement in macro F1-score), although it decreases precision. Lemmatization and stemming seem to be equally effective. In all subsequent experiments, we use lemmatization.

Table 2: Performance with documents-per-descriptor cut-off

Cut-off	Micro P	Micro R	Micro F1	Macro F1
2	80.7	58.8	68.0	47.8
3	80.6	59.5	68.4	50.0
4	80.6	60.2	68.9	52.2
5	80.6	60.9	69.4	54.1
6	80.6	61.5	69.8	55.7
7	80.6	61.9	70.0	56.4
8	80.7	62.3	70.3	57.3
9	80.9	62.8	70.7	58.7
10	81.1	63.3	71.9	59.5

As noted earlier, EuroVoc classification is known to suffer from class sparsity. To account for this, Steinberger et al. (2012) discard the descriptors that were assigned less than four times in JRC-Acquis. To gain an insight into how class sparsity affects the performance on NN13205 collection, we also discard the rarely used descriptors and re-train the model. We experiment with a cut-off threshold ranging from 2 (descriptor has to be assigned to at least two documents) to 10 (descriptors has to be assigned to at least ten documents). The results are shown in Table 2. The recall increases proportionally to the cut-off threshold, while precision increases only marginally. When only the descriptors assigned to ten or more documents are considered, micro recall improves by 6.5 percent points, resulting in a relative improvement of macro F1-score of almost 25%.

It is perhaps interesting to compare our results to that of Steinberger et al. (2012), obtained on the JRC-Acquis corpus. Steinberger et al. use a documents-per-descriptor cut-off of 4, but always assign six descriptors per document, while we assign descriptors independently of the other descriptors, based on the classifier output and the threshold. As they computed a non-standard variant of the F1-score,[7] we computed the modified F1-score in the same way for the sake of comparison. The modified F1-score on JRC-Acquis varies from 44.2% to 54.4% depending on the language. The modified F1-score at NN13205 with a cut-off of four is 60.8%. Note, however, that this comparison is for indicative purposes only, as the collections are different.

4.5. Addressing Class Sparsity

Discarding rarely used descriptors does not really address the issue of class sparsity but rather avoids it. The question arises how to practically address this issue. An intuitive approach is to rely on the hierarchical nature of the EuroVoc thesaurus. We experiment with three such techniques.

Descriptor lifting. The first technique is simply to lift the descriptors up the taxonomy tree. We replace all descriptor assignments with the corresponding microthesauri or fields, i.e., we effectively lift the descriptors to the second or first level of the EuroVoc thesaurus. The results are shown in Table 3. Expectedly, lifting to level 2 substantially

[7] We base this assessment on the analysis of JEX source code.

Table 3: Performance with descriptors lifted to thesaurus level 2 (microthesauri) and level 1 (fields)

Level	Micro P	Micro R	Micro F1	Macro F1
2	80.1	65.6	72.1	62.6
1	82.2	73.0	77.3	72.7

improves the recall (cf. Table 1), while precision remains unaffected, suggesting that most false positive assignments occur within microtheasuri. Lifting to level 1 improves recall by another 8 percent points and slightly improves the precision.

While it is obvious that this technique oversimplifies the original problem, it nonetheless does have a practical value. In the context of semi-automated indexing, one is typically aiming at automatically retrieving all plausible descriptor candidates, leaving to the human indexer the task of choosing the correct ones. In such a setting, identifying the correct field or microthesaurus might be useful for narrowing down the search. Other applications could also benefit from such coarse-grained EuroVoc classification, such as faceted search, in which the retrieved documents could be grouped based on fields or microthesauri.

Descriptor expansion. The other technique we tried out to combat class sparsity transforms the training set in a way that incorporates information stored in the descriptor hierarchy. The intuition is that the probability mass assigned to every node in the class hierarchy can be redistributed (smoothened) to cover some classes not present in the training set. We experimented with two schemes, both of which add descriptors to the original document collection: *upward expansion* (adding parent descriptors all the way up to the third level of the taxonomy) and *downward expansion* (adding child descriptors to the immediately lower level). Note that, since we work with taxonomic relations, upward expansion introduces no noise, while downward information does. In the latter case, the intuition behind descriptor expansion is that human indexers are not always consistent when deciding whether a parent or a child class should be selected, thus adding new descriptors with smaller weights to documents in the training set models this uncertainty in a simple way. The decision whether to apply expansion on a descriptor is done at the level of the whole collection, by optimizing the F1-score of that descriptor on the validation set (within the nested cross-validation loop, cf. Section 4.2.).

The classification results with descriptor expansion techniques are shown in Table 4. Upward expansion leads to slight improvements in performance (cf. Table 1), while downward expansion decreases the performance.

Recall optimization. The last technique we considered is to optimize the threshold of each model to maximize the recall. As the above experiments have shown, low recall can be traced down to low performance on sparse classes. By inverse logic then, we hope to address the problem of class sparsity by directly optimizing the recall. To this end, we again optimize the threshold of each individual model using the SCutFBR.1 method proposed by Yang (2001), only this time we optimize the F2-score instead of F1-score. The F2-score weights recall twice as much as precision. The results are shown in Table 5, alongside the previous results with F1-score optimization. F2-score optimization improves the recall by almost 5 percent points, however it decreases the precision by over 10 percent points. Overall, the macro F2-score gets improved by 1.1 percent points.

Table 4: Performance with descriptor expansion techniques

Expansion	Micro P	Micro R	Micro F1	Macro F1
Upward	79.6	60.2	68.6	48.0
Downward	72.7	57.2	64.0	43.8

Table 5: Performance with F2 (recall) optimization

Objective	Micro P	Micro R	Macro F1	Macro F2
F1	80.7	58.8	47.8	48.0
F2	70.1	63.6	47.6	49.1

5. Conclusion

We have described the work on multi-label classification of Croatian legislative documents with the descriptors from EuroVoc. We presented NN13205, a manually indexed document collection of Croatian legislative documents, which is now freely available. We performed several multi-label classification experiments on this collection. We considered several techniques to address the class sparsity problem. In particular, using upward expansion of descriptors we were able to improve the performance of the classifier, reaching 79.6% precision, 60.2% recall, and 68.6% micro F1-score.

There are a number of interesting directions for future work. First, it would be useful to obtain an estimate of the inter-annotator agreement on the NN13205. From a machine learning perspective, it would be interesting to consider multi-label classification models, hierarchical classification models, as well as combinations thereof, such as the HOMER algorithm proposed by Tsoumakas et al. (2008). Evaluation that takes into account multiple labels and hierarchy could also be considered. Finally, an interesting direction for future work are the methods for improving of annotation quality based on semi-supervised active learning, perhaps along the lines of (Settles, 2011) and (Raghavan and Allan, 2007).

Acknowledgments. We thank the Croatian Information-Documentation Referral Agency (HIDRA), now known as the Digital Information Documentation Office of the Government of the Republic of Croatia, for their support and for allowing to make NN13205 collection publicly available. Special thanks go to Maja Cvitaš and Neda Erceg for their assistance and advice. We also thank all the participants involved in the CADIAL project.

6. References

G. Boella, L. Di Caro, L. Lesmo, D. Rispoli, and L. Robaldo. 2012. Multi-label classification of legislative text into eurovoc. In *JURIX*, pages 21–30.

R. Buchan. 1983. Computer aided indexing at NASA. *The Reference Librarian*, 7(18):269–277.

N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. 2006. Hierarchical classification: combining Bayes with SVM. In *Proceedings of the 23rd international conference on Machine learning*, pages 177–184. ACM New York, NY, USA.

V. Daudaravicius. 2012. Automatic multilingual annotation of EU legislation with Eurovoc descriptors. In *EEOP2012: Exploring and Exploiting Official Publications Workshop Programme*, pages 14–20.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. 2008. LIBLINEAR: a library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

J. Hradilova. 1995. Thesaurus EUROVOC-Indexing language of the European Union. *Infoc*, 1(3):66–69.

ISO. 1985. ISO 5963-1985(E): documentation- methods for examining documents, determining their subjects, and selecting indexing terms. ISO Standards Handbook. Switzerland: International Organization for Standardization.

T. Joachims, C. Nedellec, and C. Rouveirol. 1998. Text categorization with support vector machines: learning with many relevant. In *Machine Learning: ECML-98 10th European Conference on Machine Learning, Chemnitz, Germany*, pages 137–142. Springer.

D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. 2004. RCV1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397.

N. Ljubešić, D. Boras, and O. Kubelka. 2007. Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. *Digital information and heritage/Seljan, Sanja*, pages 313–320.

M. Malenica, T. Šmuc, J. Šnajder, and B. Dalbelo Bašić. 2008. Language morphology offset: Text classification on a croatian–english parallel corpus. *Information processing & management*, 44(1):325–339.

E. L. Mencía and J. Fürnkranz. 2010. Efficient multilabel classification algorithms for large-scale problems in the legal domain. In *Semantic Processing of Legal Texts*, pages 192–215. Springer.

E. Mohamed, M. Ehrmann, M. Turchi, and R. Steinberger. 2012. Multi-label eurovoc classification for eastern and southern eu languages. *Multilingual Processing in Eastern and Southern EU languages-Low-resourced Technologies and Translation*, pages 370–394.

A. Montejo Ráez, L. Urena-Lopez, and R. Steinberger. 2006. Automatic Text Categorization of Documents in the High Energy Physics Domain. Technical report, Granada Univ. Granada.

H. Raghavan and J. Allan. 2007. An interactive algorithm for asking and incorporating feature feedback into support vector machines. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 79–86. ACM.

T. Rose, M. Stevenson, and M. Whitehead. 2002. The Reuters corpus volume 1 – from yesterday's news to tomorrow's language resources. In *LREC*, volume 2, pages 827–832.

B. Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478. Association for Computational Linguistics.

J. Šnajder, B. Dalbelo Bašić, and M. Tadić. 2008. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing and Management*, 44(5):1720–1731.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.

R. Steinberger, M. Ebrahim, and M. Turchi. 2012. JRC EuroVoc Indexer JEX – a freely available multi-label categorisation tool. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pages 798–805.

A. Sun and E.-P. Lim. 2001. Hierarchical text classification and evaluation. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 521–528. IEEE.

M. Tadić, B. Dalbelo Bašić, and M.-F. Moens. 2009. Computer-aided document indexing accessing legislation: A joint venture of Flanders and Croatia. In *Technologies for the Processing and Retrieval of Semi-Structured Documents*. Croatian Language Technologies Society.

G. Tsoumakas and I. Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

G. Tsoumakas, I. Katakis, and I. Vlahavas. 2008. Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, pages 30–44.

Y. Yang and J. O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420.

Y. Yang. 2001. A study of thresholding strategies for text categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 137–145. ACM.

M. Zhang and Z. Zhou. 2013. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 31:1.

Behaviour of Collocations in the Language of Legal Subdomains

F. Cvrček, K. Pala, P. Rychlý
Faculty of Law, University of West Bohemia; Faculty of Informatics, Masaryk University
f.cvrcek@worldonline.cz,{pala, pary}@fi.muni.cz

Abstract

In the paper we examine the collocational behaviour of multiword expression in legal sublanguages, i.e. in texts of statutory law, texts of case laws of Supreme Courts and law textbooks. We show that the comparison of collocations coming from the individual types of legal texts provides quantifiable data, which contain information about terminological nature of the observed language expressions. From the observations we made it follows that the legal language of the primary regulations considerably differs from the sublanguage of the secondary regulations. The quantitative analysis of the Czech legal texts has convincingly shown that the corpus analysis working with relatively simple means indicates the high number of changes in the texts of law regulations. In this way the changes also show that the corpus analysis also reflects the problems in our society – too many and fast changes in the legal texts prevent lawyers from the correct handling the individual court cases. In the paper we also exploit the results of the project PES (Právní elektronický slovník, Legal Electronic Dictionary.

Keywords: collocations, corpora, legal subdomain, legal electronic dictionary

1. Introduction

In the paper we explore behaviour of collocations in legal subdomains (sublanguages), i.e. in texts of statutory law, law cases of the Supreme Courts and law textbooks. The comparison of collocations from the individual types of the legal texts provides quantifiable data which give information about terminological nature of the observed collocations. From the performed measurements it follows that the legal language of the primary regulations considerably differs from the sublanguage of the secondary regulations.
Linking corpora and dictionary allows us to test the new regulations from the point of view of new terminology (every new word) and the integration of a new regulation into the existing legislation. This provides a significant support for legislative work. At the same time the legal language and general (common) language can be compared.

2. Resources

As the material for our research we have used the following resources (corpora): CzLaw with size 20,643,133 tokens which is divided into two subcorpora:

- Primarnipredpisy (Primary Regulations) (Ústava a platné zákony ČR – The Constitution laws and laws in force of CR) with 12,249,408 tokens,
- Sekundarnipredpisy (Secondary Regulations) (vyhlášky, nařízení – government decrees and decrees issued by Ministeries in force) with 8,393,725 tokens.

The size of these corpora is not large but it is sufficient (complete collection of statutory law in force from 1989) for demonstrating the basic collocational tendencies we are interested in as they can be examined by means of Word Sketches (Kilgarriff et al, 2004). For the contrastive analysis we have also used the corpus CzechParl comprising 51.4 mil. tokens and consisting of the transcribed recordings of the speeches made by MPs in the Parliament of Czech Republic. The texts are of legal nature and their size is sufficient for our purposes.

The corpus of the primary regulations contains what is produced by the Parliament and secondary regulations – what is produced by the government. Presently, we are concerned with the state of the valid regulations by the end of 2012. Each year, a new corpus of the regulations valid by the end of the year is created, which will allow us to make a comparison.

3. Czech Legal Electronic Dictionary (PES)

In the paper we also exploit the results of the project PES (Právní elektronický slovník, Legal Electronic Dictionary, see http://deb.fi.muni.cz/pes (Cvrček, Pala, Rychlý et al, 2012), in which corpus linguistics and juristic approaches are successfully intertwined. The PES represents an analysis of the legal terminology based on the language of basic law textbooks covering the individual law branches, on the language of the law acts (corpus of the valid law acts including the Constitution of Czech Republic), on the language of the secondary regulations (corpus of the rules and regulations on the central level of Czech Republic), on the language of the law cases (corpus of the judicial decisions made by the High Courts from 1990), and also on the general written Czech language (corpus Czes2 with 465,102,710 tokens).
The system PES (a collection of databases, corpora and programs) makes it possible to investigate the legal language and its changes. Thanks to its size, it practically captures the whole Czech law system, we, in fact, obtain the full picture of the law complexity on the linguistic level for the first time. Software system PES is regularly updated and can be accessed by the all users who would like to use it for the research and teaching if they apply for it to Dr. F. Cvrček from the Institute of Law and Government, Czech Academy of Sciences.
It should be added that the PES is not just a dictionary, but together with the dictionary and legal ontology it contains corpora of the primary legislation, secondary legislation and general language. Legal data comes from databases which have been created in ÚSP (Institute of Government and Law, Czech Academy of Sciences) since 1985 and con-

tains 100,000 documents representing the legal system at the level of law regulations and case law.

4. Tools Used for the Analysis

The individual mentioned corpora we handled using the corpus manager Manatee/Bonito (Rychlý, 2007) with the built in module for processing Word Sketches (Kilgarriff et al, 2004). The manager Manatee/Bonito allows users to search the mentioned corpora, to obtain the concordances from them, to observe the frequencies of the individual expressions (legal terms) and especially to observe their collocational behaviour, to get key words and compare the individual subcorpora on various levels.

4.1. The Analysis

The first evidence about the difference of primary and secondary sublanguages of the legal regulations is provided by the comparison of the key words in the both subcorpora. The lists of the key words have been created for both subcorpora by means of comparison word frequencies with the reference corpus czTenTen (Jakubíček et al, 2013). The czTenTen corpus is a web corpus, it was created by crawling the Web by the web crawler SpiderLing (Suchomel, 2012) during April 2012. It was cleaned (headers, menus etc. removed) and any duplication and near-duplications inside or between pages was removed. As a result of the building process, the corpus has very wide coverage of topics, it is a generic corpus and it plays a very good role as a reference corpus. Various parameters of the used corpora (and subcorpora) are summarised in the Table 1.

During computing we have used what is called the Average Reduce Frequency (Savicky at al, 2003), which automatically filters words occurring in one or just few documents. We have compiled several lists with the various length containing always statistically most significant key words from the given subcorpus. The comparison of the respective lists shows that only 60 % words are common in the corresponding lists. An example of key-word lists comparison is in the Table 2, it compares top 30 key-words in both subcorpora. The common key-words are very technical: paragraph labels, labels of different sections of documents (paragraphs, parts, regulations etc.) and document labels (number, law, etc), there is only one verb: *stanovit (to state)*. There are three verbs in Primarnipredpisy which do not occur much in Sekundarnipredpisy: *nahrazovat (to change), vkládat (to insert), zrušovat (to remove)*, all are results of constant changes in Czech law system.

In a similar way we have compared collocations of the individual key words common to the both mentioned legal subcorpora. The collocation lists have been created using the Sketch Engine (Kilgarriff et al, 2004) providing statistically significant collocations based on the respective grammatical relations. Word Sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. The system, in addition to using a well-founded salience statistics and lemmatization, finds collocations by using grammar patterns. Rather than looking at an arbitrary window of text around the headword, it looks, in turn, for each grammatical relation that a word participates in. The grammar patterns are based on a powerful query language and they use regular expressions over part-of-speech tags. An example: if we wish to define the English verb-object relation, we first note that, lexicographically, the noun we wish to capture is the head of the object noun phrase, and that this is generally the last noun of a sequence that may include determiners (DET), numbers (NUM), adjectives (ADJ) and other nouns (N). We also note that the object noun phrase is, by default, directly after the verb in active sentences, and that the lexical verb (V) is generally the last verb of the verb group. Adverbs (ADV) may intervene between verb and object. Taken together, these give a first pass definition for a "verb-object" pair, as "a verb and the last noun in any intervening sequence of adverbs, determiners, numbers, adjectives and nouns". In the Sketch Engine formalism, using the tags given in brackets above, this is

```
1:"V" "(DET|NUM|ADJ|ADV|N)"* 2:"N"
```

The *1:* and *2:* mark the words to be extracted as the first and second arguments of the grammatical relation. The above example defines the English verb-object relation, similar rules have been defined for Czech language. As a result, the collocations are not simple co-occurrences in a predefined window of words, rather it is a result of the special type of shallow parsing.

The repeated comparison of the lists between the two subcorpora shows that for some words the portion of common collocations is smaller than 30 %. As an example we can offer the word *territory (území)*, for which from 37 grammatical relations only 18 ones have at least one relation common to both subcorpora and only 6 relations have more than three common collocations. The example of visual comparison for the the subcorpora is in Figure 1.

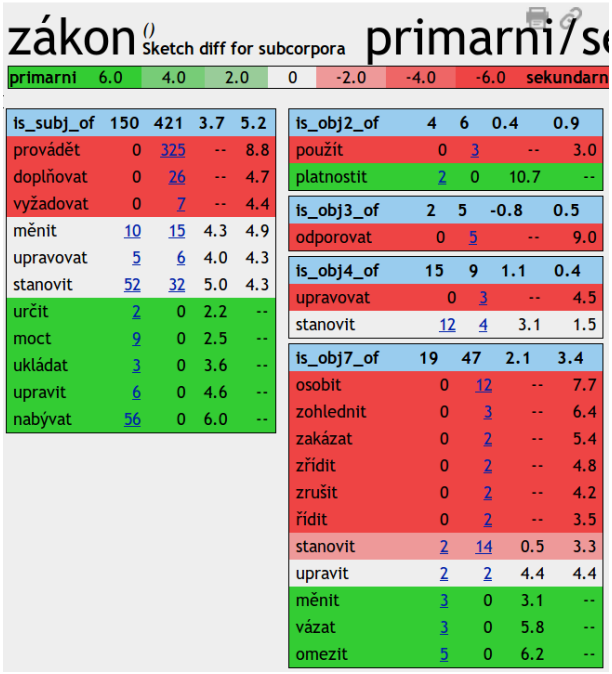


Figure 1: Visual comarison of collocations in subcorpora.

	CzLaw	Primarnipredpisy	Sekundarnipredpisy	czTenTen
Tokens	206,43,133	12,249,408	8,393,725	5,214,920,358
Sentences	318,184	160,071	158,113	291,413,460
Documents	5,294	2,009	3285	9,296,371

Table 1: Corpora sizes

Only in Primarnipredpisy	Only in Sekundarnipredpisy	Common key-words
nahrazovat (replace)	g	b
orgán (body)	nabývat (acquire)	c
povinen (obliged)	nařízení (decree)	d
povinnost (obligation)	podle (according to)	e
písmeno (letter, point)	příloha (annex)	f
vkládat (insert)	v. (see)	odst. (par.)
znít (have a wording)	vyhláška (regulation)	odstavec (paragraph)
zrušovat (cancel)	zařízení (institution)	osoba (person)
zvláštní (special)	údaj (information, data)	právní (legal)
úřad (office)	účinnost (efficiency)	písm. (letter, point)
		předpis (directive)
		příslušný (respective)
		sb. (collection of acts)
		stanovený (set up)
		stanovit (to set up)
		ustanovení (statute)
		uvedený (mentioned)
		znění (wording)
		zákon (law act)
		č. (number)

Table 2: Comparison of keyp-words lists

At the first glance it would seem that laws on one hand and regulations and rules on the other belong to the same language, the exact statistical analysis shows, however, that they represent two considerably different domains, which speak by the distinct sublanguages.

5. What Follows from the Collocational Analysis of Legal Texts

A remark has to be made which may be considered unusual in this sort of text: the semantic analysis of the legal texts from our corpora indicates that there are some social and political problems which have immediate and unpleasant consequences for Czech society.

Observe that the most frequent word in the corpus of the legal texts is *zákon (law act)* and its simple Word Sketch shows that the most frequent collocations with genitive case are *znění zákona (wording of law act), změna zákona (change of law act), návrh zákona (proposal of law act), doplnění zákona (amendment of law act)*. When noun *zákon (law act)* collocates with verbs the most frequent ones are *měnit zákon (to change law), stanovit zákon (to set a law down), doplňovat zákon (to supplement to a law)*, etc. This provides clear evidence that in Czech legal system we face an excessive and abnormal revising of law acts which, in fact, endangers the legal system of Czech Republic as such. The further detailed quantitative analysis of the Czech legal texts convincingly shows that the linguistic research working with relatively simple means confirms the indicated situation, i.e. the existence of the jeopardy consisting in low transparency and incomprehensibility of the links between law regulations. All this threatens seriously a standard exploitation of law in Czech Republic. We would like to stress that the results of our research provide practical results that will be presented to the politicians to become aware what they are really doing.

The analysis of the most frequent word of legal texts, ie „the Law Act" indicates the main problem of the Czech legislation – the inflation of modifications. This hypothesis was verified on legal databases based on the number of modifications (hyperlinks, amended points, etc.) since 1918 (see Figure 2). Furthermore, based on typical words in the titles of regulations (such as to change, add to, change, etc.) the ratio of modifications was compared to the ratio of the total production regulations for the V4 countries and Austria (see Figure 3 and Figure 4). It turned out that the Czech Republic is the worst and is followed only by Slovakia. Thus corpora can be used to quickly indicate legal problems. At the same time on the ground of the corpus of general language it can be shown how the problem is perceived by society. For the primary regulations a significant indicator is the relation *preceding-verb–verb* (change, supplement) – 60 %, for secondary regulations *preceding-verb–verb* – 90 %. For the corpus of general language relation with genitive shows 2–40 % (change, supplement,

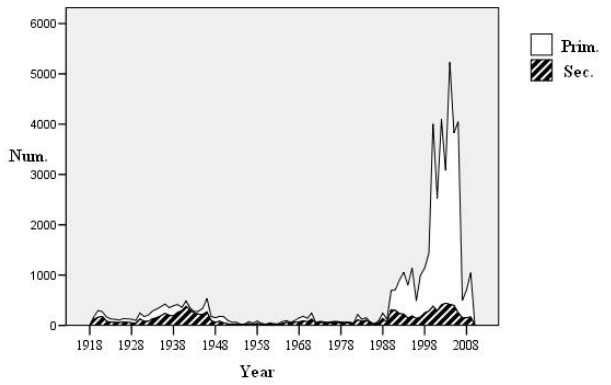


Figure 2: Number of modifications, primary and secondary legislation – Czech Republic

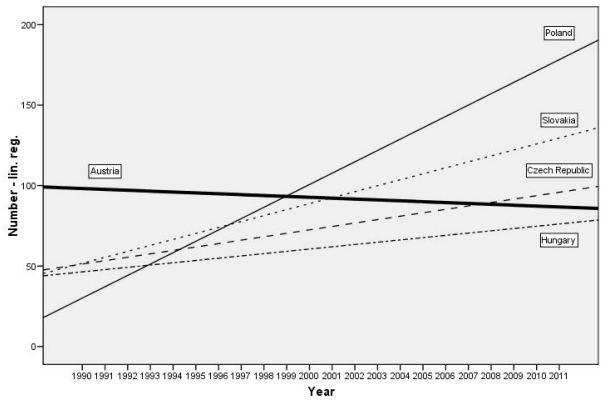


Figure 3: Modifications of primary legislation from year 1990 – CZ, SR, PL, HU, AU

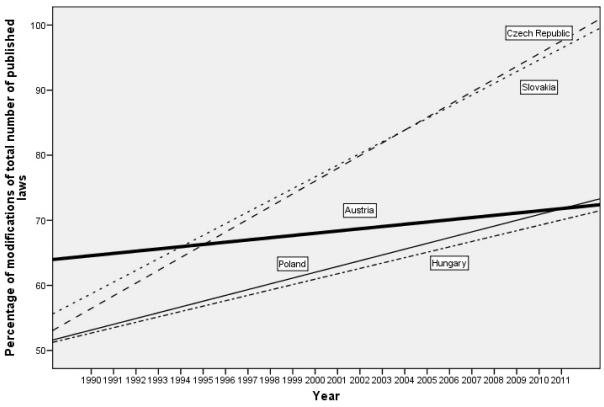


Figure 4: Modifications of primary legislation from year 1990 expressed as percentage – CZ, SR, PL, HU, AU – trends

amendment). The expression *amendment* does not occur in the legal corpora, but in the general language base it is a term for denoting frequent changes.

6. Conclusions

In the paper we have offered a collocational view and comparison of the selected legal subdomains and their sublanguages. For this analysis we have used the tool called Sketch Engine as well as corpus manager Manatee/Bonito. According to our knowledge this is the first case when the Sketch Engine has been used for the collocational analysis of the legal language with interesting results. It can be seen that Word Sketches provide quite a detailed explanation of the multiword expressions in the legal texts. The comparison of Primary and Secondary Regulations subcorpora shows that these subcorpora form quite different language variants in the field of legal texts.

These findings are also supported by the data contained in the Czech Electronic Legal Dictionary (PES), which is a result of the larger project dealing with Czech legal terminology and its hierarchical (ontological) structure. In the PES the relations between legal terms and the concepts they are expressing are well reflected as well as the changes caused by the frequent amendments of the individual laws.

The theoretical linguistic analysis thus provides evidence that the obtained results have concrete practical consequences – they show that the Czech law system loses transparency necessary for its reliable functioning.

7. Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

8. References

Cvrček, F., Pala, K., Rychlý, P. et al. (2012). PES (Právní elektronický slovník – Electronic Legal Dictionary), its web page: http://deb.fi.muni.cz/pes.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P and Suchomel, V. (2013). The TenTen Corpus Family, *7th International Corpus Linguistics Conference*, Lancaster.

Kilgarriff, A., Rychlý, P., Smrž, P. and Tugwell, D. (2004). The Sketch Engine *Proceeding of Euralex.* Lorient, France, July, 105–116.

Mráková, E. and Pala, K. (2010). Legal Terms and Word Sketches: a Case Study, *Proceedings of the RASLAN Workshop*, Karlova Studánka, 31–40.

Pala, K., Rychlý, P. and Šmerk, P. (2012). Automatic Identication of Legal Terms in Czech Law Texts, *Semantic Processing of Legal Texts*, Springer, LNAI 6036, 83–94.

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager, *Proceedings of the RASLAN Workshop*, Karlova Studánka, 65–70.

Savicky P. and Hlavacova J. (2003). Measures of Word Commonness. *Journal of Quantitative Linguistics*, Vol. 9, No. 3, pp. 215–231.

Suchomel, V. and Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora, *ACL SIGWAC Web as Corpus*. Lyon.

State of the ART: an Argument Reconstruction Tool

Radboud Winkels, Jochem Douw, Sara Veldhoen
Leibniz Center for Law,
University of Amsterdam
winkels@uva.nl

Abstract

This paper describes the outcomes of a series of experiments in automated support for users that try to find and analyse arguments in natural language texts in the context of the FP7 project IMPACT. Manual extraction of arguments is a non-trivial task and requires extensive training and expertise. We investigated several possibilities to support this process by using natural language processing (NLP), from classifying pieces of text as either argumentative or non-argumentative to clustering answers to policy green paper questions in the hope that these clusters would contain similar arguments. Results are diverse, but also show that we cannot come a long way without an extensive pre-tagged corpus.

Keywords: argument mining, clustering, cluster tendency, policy modelling, machine learning

1. Introduction

Before publishing a policy white paper, the European Union often publishes a draft, a green paper, to stimulate discussion and enable public consultation. The green paper provides the opportunity to companies and individuals to respond to the draft and provide arguments in favour or against it. Typically such a green paper raises issues and ask questions like "Should there be encouragement or guidelines for contractual arrangements between right holders and users for the implementation of copyright exceptions?".[1] Exploring and indexing these replies and their arguments from external sources is difficult and time consuming. The goal of EU FP7 project IMPACT ("Integrated Method for Policy Making Using Argument Modelling and Computer Assisted Text Analysis") was to provide means to support this process.[2] This includes a so-called "Argument Reconstruction Tool" (ART) that enables users to easily copy and store text fragments and relate them using formal argument structures. Part of the foreseen functionality of the tool was to help the user by finding text fragments that contain arguments and possibly suggesting argument schemes that are used.
This paper introduces the ART and focusses on two experiments in automated argument finding and reconstructing.

2. The ART

The ART is implemented as a Rich Internet Application (RIA). Arguments are stored using a separate storage class that abstracts away from the current MySQL implementation. Users can copy and paste any piece of text into the system by hand and construct arguments at different levels of detail:

Unary Relations (UR) We are enabling users to start with annotating texts with qualifications like "there is an argument somewhere here" or "this is a proposition that is part of an argument". These are unary relations on the pieces of text, usually consisting of one or more sentences.

Binary Relations (BR) In addition to that we enable users to make binary relations between arguments or parts of arguments. These binary relations can e.g. be of the form "A supports B" or "A attacks B". These relationships can actually exist on several different levels: it can e.g. be a relation between two entire arguments (represented in either of the three states below) or between two variables (necessitating the argument to be modelled at the PCLAS level).

Abstract Argument Scheme (AAS) This relationship connects one or more premises to a conclusion.

Proposition Level Argument Scheme (PPLAS) We make a distinction between different sorts of premises based on an argument scheme. For the Argument from Credible Source (ACS) scheme, we could make a distinction between the atomic terms "Newton was an expert in science", "Newton said that things always fall down" and "Statements about things falling down fall within the domain of science". These statements have three types, that could be called respectively "Credible source assumption", "Person asserts statement", and "Asserted statement within domain".

Predicate Level Argument Scheme (PCLAS) This is the finest level of argumentation representation. When we take the ACS scheme as example again, we make a distinction between atomic statement types "expert", "statement" and "domain". These have a fixed meaning within the ACS scheme, but can also be coupled as predicates by saying asserts(*expert*, *statement*), or at the instantiated level asserts("Einstein", "All things fall down").

All these schemes can be either uninstantiated or instantiated.
The ART currently has three basic argumentation schemes:

1. General Argument Scheme: The most simple one, just consisting of one or more premises and a conclusion.

[1] From "Copyright in the Knowledge Economy".

[2] See `http://www.policy-impact.eu/` for more information.

2. Credible Source Argument Scheme: It consists of a proposition from a certain domain stated by a particular source. See figure 1.
3. Practical Reasoning Argument Scheme: Consists of an action proposed by an agent in particular circumstances described by one or more propositions, leading to consequences described by one or more propositions to promote one or more values.

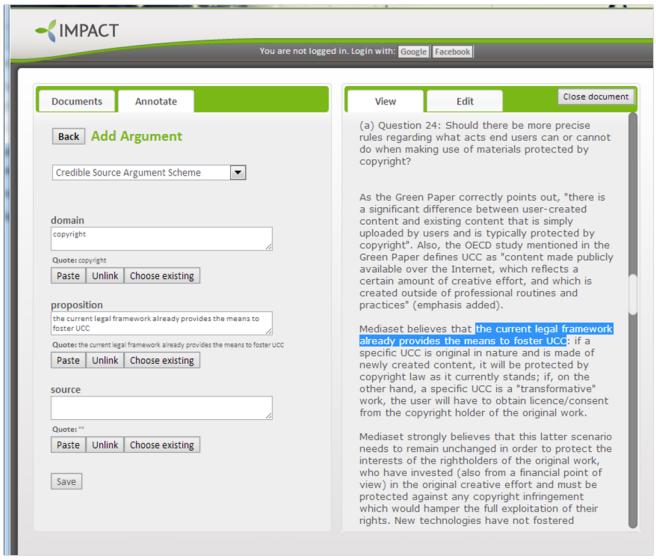


Figure 1: *A partially filled credible source argument scheme.*

3. Extraction of Arguments

Manual extraction of arguments from a text is a non-trivial task. In (Mochales and Moens, 2011), an example is given of three annotators that had to identify arguments in verdicts of the ECHR.[3] They write: *"The overall process took more than a year and included three annotators and one judge to solve disagreements. Once the task was completed, the annotation obtained a 75% agreement between annotators [...]"* . It would be helpful if the machine could detect the use of arguments and suggest schemes and perhaps even prefill them and present them for verification to the human users.

3.1. Related Research

In the ART, arguments can be extracted manually by users. We have the ambition to employ natural language processing (NLP) to recognise the arguments inside a natural language text such as a green paper, a website or a blog. In general one can state that up to the beginning of the IMPACT project in 2010, hardly any research had been devoted to automated argument reconstruction from natural language texts (cf. (Moens et al., 2007), (Palau and Moens, 2009)).

Brüninghaus and Ashley (Brüninghaus and Ashley, 2005) built systems that recognise relevant factors in legal texts and then proceed to generate (and evaluate) an argumentation from those facts. Classifiers were made to determine if a certain factor appeared in each sentence. These factors came from a list of factors from trade secret cases, and are more specific than the arguments that a generic tool should recognise. Different machine learning approaches were tested to train these classifiers, with three different forms of data representation. TiMBL worked best with data represented as propositional patterns (F-measure of 0.26).

An actual attempt at argument detection has been made by Mochales Palau and Moens (Palau and Moens, 2009). They perform three steps: 1. classification of a proposition as argumentative or non-argumentative; 2. classification of an argumentative proposition as a premise or a conclusion; 3. detecting the argument structure.

In a corpus based on diverse sources (the so-called structured Araucaria corpus consisting of 641 arguments from newspaper articles, online discussion boards, and magazines) they were able to detect arguments with 73% accuracy; classify premises and conclusions with a F1 measure of about 70%, and detect argumentation structures with about 60% accuracy. The argument structure is detected using a context-free grammar. The classification was attempted with both machine learning classifiers and context-free grammars, with the machine learning classifiers (maximum entropy model and support vector machines) leading to the best results.

A somewhat different approach is to start with a classification of the relation between two text fragments rather than the classification of the text fragments themselves. Marcu and Echihabi (Marcu and Echihabi, 2002) focus on the automated recognition of discourse relations, which are descriptions of how two spans of texts relate to each other. They created a corpus containing different text fragments and the relation between them, confining themselves to the relations contrast, cause explanation-evidence, elaboration and condition. They then used Naive Bayes classifiers to distinguish between two relations, which had a performance of between 64% and 93%, depending on the relations that were compared.

These approaches suggest that a machine learning approach will be better for the task of detecting arguments than a pattern-based approach, but that identifying relevant patterns is still valuable, as they can be included as features for the machine learning approach.

4. First Experiment

As explained above, literature suggests the use of machine learning techniques. However, the dataset required to train such machine learning techniques will be developed using the ART tool once it is operational. Unfortunately we were not able to accumulate a large enough dataset from other sources, so we resorted to keyword-based tagging based on manual inspection of sources.

The domain consists of replies to the EU green paper "Consultation on the Commission Report on the enforcement of intellectual property rights".[4] These documents are mostly written in a neutral style, with a low amount of sentiment cues. The arguments provided often consist of just propositions without keywords indicating their role or the fact that

[3] The European Court of Human Rights in Strasbourg, France.

[4] The replies can be found at : `http://ec.europa.eu/`

it is an argument at all. Domain knowledge and common sense are required to reconstruct the argumentation in these responses. Finally, almost every argument is an implicit "argument from position to know" (Walton, 2002). This is inherent to the context of green paper discussions, which is that companies and organisations establish themselves as being in the position to know about the topic at hand and then try to convince the EU of a particular standpoint.

4.1. Keywords and Regular Expressions

The first step was to see if the documents contained any keywords that indicate the use of argumentation. The following documents were used as training set.

Source	Total words
ANBPPI/BNVBIE	5165
Google	4830
Bits of Freedom	2150
Ericsson	1919
Business Europe	1068

Three observations can be made. (1) The frequency of most keywords, if not all, is very low (a small portion is shown in table 1). The documents contain arguments in nearly every paragraph, but only a small portion of these arguments uses identifiable keywords. (2) The use of argumentative expressions, linguistic constructions and vocabulary differs dramatically *over* documents, but is rather consistent *within* a document. This is one of the reasons for the overall low frequencies of keywords. (3) The keywords that were useful can be divided in roughly three categories: *Structure segments* that indicate structural relations between sentences (e.g. for example, firstly); *Argumentation segments* that indicate argumentational relations between (parts of) sentences (e.g. concludes, therefore, in contrast with, see table 1); and *Sentiment segments* that are not directly linked to argumentations but do indicate the expression of an opinion which can indirectly indicate that an argumentation is used (e.g. essential, believe). For more extensive research see (Knott and Dale, 1994).

Segment	**BoF**	**Google**	**Ericsson**	**BEurope**	**ANBPPI**	**Total**
Argumentation segments						
however	1	4	3	1	7	16
thus / therefore	2	0	4	0	6	12
lead(s) to / has resulted in / result	5	0	2	1	2	10
conclude(s) / conclusion	6	1	1	0	0	8
assumption/assume	3	1	1	0	0	5
pointed out	0	4	0	0	0	4
at odds	4	0	0	0	0	4
since	1	1	1	0	1	4

Table 1: Most frequent argumentative keywords in train set.

The next step was to construct regular expressions from these keywords to tag sentences with an argumentation indication in the *test* set. Three were created: one that matches any of the keywords or combinations of them, one that indicates some sort of conclusion and one that indicates some sort of premise. As an example we present the regular expression for conclusions below:

```
(therefore|conclu(de(s|d)?|sions?)?|in fact|
thus|hence|(this|that) is why)|
(support(s|ed|ing) the conclusion|
In sum|hereby|by doing so)
```

The regular expressions were applied on the test set consisting of the following sources:

Source	Total words
PPL UK	1181
Royal TNT Post	1264

When applying the regular expression that matches any keyword on our test set the following confusion matrix was achieved:

		Manual					
		Arg	Ntrl	Ttl	Prec	Rec	F
Tagging	Argum	16	7	23	69.6%	40.0%	50.8%
	Neutral	24	65	89	73.0%	79.5%	76.1%
	Total	40	72	112	72.3%	72.3%	72.3%

About 35% of the sentences in the test set are manually tagged as argumentative; not even half of these were found using the regular expression (recall of 40%). Only 7 sentences were incorrectly classified as argumentative (few false positives). An obvious reasons for the low recall is the observed difference in language use across authors.
When applying the other two regular expressions, both recall and precision are very low for finding conclusions (F-score of 14.3%) and low for premisses (F-score of 46.8%). Although the results are in some cases quite good, there are two factors that must be taken into account. Firstly, the size of the train and test set is too small to get real representative results. Secondly, recall and f-score values are much higher for the neutral classes than the actual classes we want to find (*Argumentative*, *Conclusion* and *Premise*). Detecting *Argumentative* works better than detecting premises, which works better than conclusions, which score the worst.

5. A second Experiment

Since we do not have a tagged corpus of arguments, neither in the domain of EU green papers, nor in any other comparable domain, we decided to explore the use of unsupervised techniques. Can we find clusterings of answers to green paper questions that correlate to the use of specific types of arguments? Even if we cannot decide which argument type is exactly used, it may help policy analysts if we can provide them with clusters of similar ones.
A different EU Green Paper on "Copyright in the Knowledge Economy" contains 25 questions belonging to five distinct topics. We have used the 159 unique replies in English (from the 374 replies in total). They contain around 1300 answers to specific questions, differing in length.
In GATE[5], we created a pipeline to annotate the questions and answers in the documents after exporting them to plain text. The output of this pipeline was a set of XML documents with the annotations as in-line XML tags. We have

[5]GATE is open source software capable of solving many text processing problems, see http://gate.ac.uk/

only taken into account the answers to single questions (as opposed to general remarks and answers to a range of questions). An answer consists of one or more lines of text. The number of the question being addressed was assigned as an attribute to the XML-tag for every answer.

5.1. Clustering

We have compared a number of clustering methods. A distinction can be made between partitioning and hierarchical approaches. Partitioning cluster algorithms output a hard partition that optimizes a clustering criterion. Hierarchical algorithms produce a nested series of partitions based on a criterion for merging or splitting clusters based on similarity (Jain et al., 1999). Applying different hierarchical clustering methods did not seem to work; we mostly got one cluster containing much (>95%) of the data. Partitioning methods resulted in more equally sized clusters, so we have focused on these algorithms.

The first method we used is Expectation Maximization (EM), which assigns a probability distribution of each instance indicating the probability of it belonging to each of the clusters. This algorithm is capable of determining the number of clusters by cross validation (Moon, 1996). Another method is SimpleKMeans. It starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers (Jain et al., 1999). XMeans and FarthestFirst are extensions of the SimpleKMeans, determining the number of clusters and choosing the initial centroids to be far apart respectively. Finally we applied sIB (Sequential Information Bottleneck), which is like K-means, but the updates aren't performed in parallel (Slonim et al., 2002).

5.2. Finding Topics

First we tried a bag-of-words approach to find clusters of documents, i.e. complete answers. All answers to all questions were taken into account. The attributes source, question number and the topic of the question were added as attributes to be used for the analysis; these were not handed to the clusterer. The text content of the answers was filtered using a stop list[6].

The data was then loaded into WEKA Explorer[7] where the content attribute was converted to a series of attributes serving as a bag-of-words. The filter StringToWordVector was used, applying IDF-TF Transform and normalizeDocLength (for normalizing the values). The minTermFreq was set to 10, thus creating around 100 attributes. The outputWordCounts was set to true, creating numeric values rather than booleans. Finally, a stemming algorithm was used to map syntactically related words to the same stem.

We applied EM clustering to the data, leaving the number of clusters to be created open. The random seed was set to 100 (default). The algorithm grouped the 1301 instances into 11 clusters, with cluster sizes ranging from 39 to 266. Every instance in the dataset is an answer to a specific question, belonging to a topic. Beside, each instance has an origin, a source document. Three matching matrices were built relating the clusters to questions, sources, and topics. The latter is shown here for illustration:

Cluster → Topic ↓	0	1	2	3	4	5	6	7	8	9	10
General	49	34	9	12	6	73	1	9	60	3	16
ELA	28	40	3	114	9	54	52	17	58	7	22
EPD	12	141	3	1	20	2	2	12	27	1	38
TR	17	36	87	2	38	18	3	16	16	26	3
UCC	61	15	2	0	2	2	2	6	11	2	1

ELA = Exceptions Libraries Archives; EPD = Exceptions for People with Disability; TR = Teaching Research; UCC = User Created Content

There are many evaluation metrics available to define the extrinsic quality of a partitioning. In (Amigó et al., 2009) a wide range of metrics is analyzed according to a few intuitive constraints. The B-Cubed metric was found to be the only one satisfying all the constraints. We have used this metric to compare the clustering to the three classifications. The precision and recall are computed for each entity in the document and then combined to produce final precision and recall numbers for the entire output.

The recall, precision and F-score of the clustering compared to the three classifications are:

Classification	Precision	Recall	F-score
Question	0.123	0.309	0.176
Topic	0.420	0.219	0.288
Source	0.027	0.232	0.049

Although the first experiment showed that linguistic constructions and vocabulary differed from writer to writer, in this experiment we see that the clustering tends to correspond more to the (topics of the) questions than to the authors: compared to the other two, the scores of the 'source' classification are quite bad. There is hardly any correspondence between the author of a reply and the cluster it is assigned to. Note that in this experiment the closed-class or function words were filtered out of the text, which was not the case in the first experiment.

This finding endorses our idea of using lexical analysis to find pieces of text expressing the same ideas or subjects. However, the scores on the other two classifications are quite low as well, so it is very well possible that there is not enough information in the bag of word features to get a proper semantic grouping.

5.3. Finding Arguments

This section describes the experiments with a finer granularity. The dataset contains all answers to a specific question, the instances are the paragraphs that the answers consist of. We aim for a clustering that expresses lines of argumentation. The procedure to represent the data is the same as before except that the minTermFreq was set to 4, because the dataset is much smaller and all terms are less frequent.

The methods EM, SimpleKMeans, XMeans, FarhestFirst and sIB were all applied to the datasets containing the answers to question 19 and question 6. EM and XMeans were run with no number of clusters specified. Furthermore, all methods were executed with the number of clusters to be created set to $2 \leq k \leq 6$. We have used EuclideanDistance as a distance function when needed. The random seed was set to 27 and 42 when this parameter was needed.

Because of the many dimensions in our data, presenting

[6] ftp://ftp.cs.cornell.edu/pub/smart/english.stop

[7] WEKA is a popular suite of machine learning software, see http://www.cs.waikato.ac.nz/ml/weka/

them in a comprehensible way is quite challenging. WEKA provides a visualization tool, which is a scatter plot containing all the instances. Even though this tool works intuitively and is capable of comparing any two dimensions, it does not give insight in the coherency of all the dimensions. Instead, we export the data to excel and use sorting and conditional formatting to visualize results. We use two methods for visualization of the clustering, one is instance based (attributes along the columns and the instances along the rows) and the other cluster based (clusters along the rows). An example of the latter can be seen in figure 2.

Analysis

Cluster evaluation metrics can be extrinsic, based on comparisons between the output of the clustering system and a *gold standard*. Since we do not have a gold standard (yet), we need to resort to intrinsic metrics. These are based on how close elements from one cluster are to each other, and how distant from elements in other clusters (Amigó et al., 2009). Furthermore, we have performed a meta-clustering to compare the clusterings of different algorithms and/or different runs of the same algorithm.

Many internal validation measures exist. We have chosen the 'index I' measure as described by (Maulik and Bandyopadhyay, 2002), which has a reasonable performance and is quite intuitive.

It is defined as:

$$I = (\frac{1}{NC} \times \frac{\sum_{x \in D} d(x,c)}{\sum_i \sum_{x \in C_i} d(x,c_i)} \times max_{i,j} d(c_i,c_j))^P$$

where D: data set; c: center of D; P: number of attributes (dimensionality) in D; NC: number of clusters; C_i: the i-th cluster; c_i: center of C_i; $d(x,y)$: (Euclidean) distance between x and y

A high I index corresponds to a good clustering. We computed this metric from 30 clusterings on the dataset 'question29': three methods $\{EM, KMeans, sIB\}$, five cluster sizes $\{2,3,4,5,6\}$, and two random seeds $\{27,42\}$. The respective values are plotted in figure 3.

Looking at figure 3, we can clearly see a correspondence between clustering quality and the number of clusters. Extrapolation of the negative correlation might even indicate that no natural partitioning exists in the data. Furthermore we see that the sIB algorithm tends to score worse than the other two. Besides, in some cases the random seed has quite some influence on the scores.

The I index provides means to compare different clusterings on the same dataset. We can use it to decide which clustering best matches the natural partitioning in the data. We can also use this technique for determining the proper number of clusters to aim for. But beside this, it doesn't tell us much about the nature of the data itself. The scores can be interpreted in relation to each other, but do not give an absolute measure.

On a higher level, we can compare the clusterings of different algorithms and/or different runs of the same algorithm. We are interested in deriving a consensus solution, presuming that if many clustering algorithms reveal the same structure, there must be some intrinsic partitioning in the data. This method is loosely based on the idea of Cluster Ensemble (Strehl and Ghosh, 2003). The technique we have used for this investigation is meta-clustering: we have run an EM clusterer with 13 clusterings (partitionings) as attributes (features). With the number of clusters unspecified, 9 clusters were created. We have also run the EM algorithm with the number of clusters set to 2 and 5. The resulting partitionings were unstable as well, which strengthens our belief that no partitioning can be found.

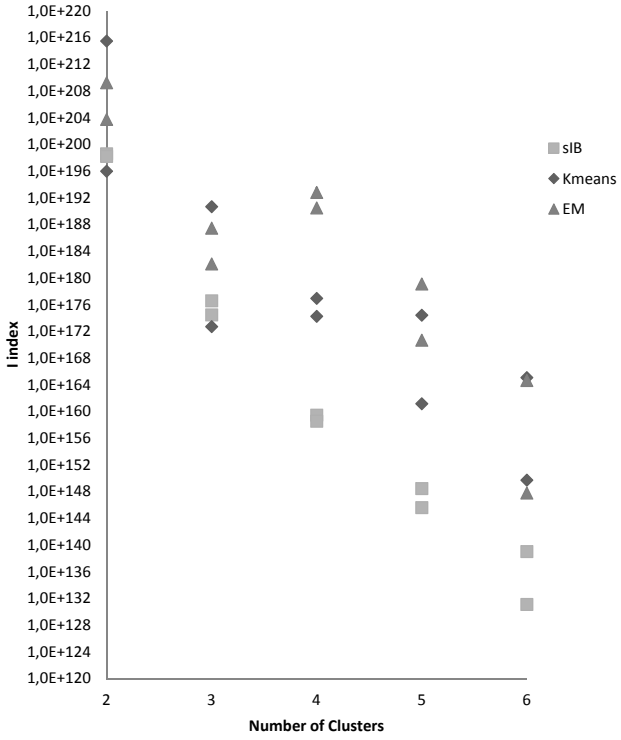


Figure 3: *I indices for 30 clusterings*

Cluster Tendency Although we did not find any indication of a natural grouping, the absence of it is hard to prove as we might have used the wrong technique or applied the wrong settings. The I index defines the quality of a clustering. Our objective is not to reveal the best possible clustering in the data however, but to investigate whether any clustering exist. "All clustering algorithms will, when presented with data, produce clusters - regardless of whether the data contain clusters or not. The first facet of a clustering procedure is actually an assessment of the data domain rather than the clustering algorithm itself. This is the field of *cluster tendency*, unfortunately this research area is relatively inactive" (Jain et al., 1999).

One method for assessing the cluster tendency of a set of objects is called VAT (Visual Assessment of (cluster) Tendency) (Bezdek and Hathaway,). First a distance matrix is created with the instances along both the axes, thus providing a pairwise (two-dimensional) interpretation of high-dimensional data. Secondly the instances are reordered according to an algorithm that is similar to Prim's algorithm for finding a *minimal spanning tree* of a weighed graph. Both matrices can then be displayed as *dissimilarity images*. The pairwise dissimilarity of the objects (the value in the distance matrix) determines the intensity or gray level of the corresponding pixel in the image. Clusters are indi-

Cluster:	size:	39-bag-imb	43-bag-retain	40-bag-incentiv	11-bag-negot	46-bag-tool	27-bag-schol	41-bag-leg	36-bag-reus	1-bag-ca	42-bag-part	38-bag-governm	44-bag-right	33-bag-level	26-bag-pol	13-bag-author	28-bag-common	84-bag-agr	88-bag-fund	45-bag-teach	2-bag-publ	24-bag-open	12-bag-research	57-bag-encour	58-bag-goal	79-bag-copyright	9-bag-purp
Cluster0	0,27	0	0	0	0,4	0	0	0,1	0,1	0,4	0,3	0	0,3	0	0	0,1	0,1	0,4	0,1	0,5	0,4	0	0,5	0	0,1	0,3	0,6
Cluster1	0,01		4,5	4,2	4,1	3,7	3,8	3,4	3,4	3,3	3,1	3,1	2,6	2,3	2,1	2,1	1,7	1,6	1,6	1,5	1,7	0	1,3	0	0	1,1	1
Cluster2	0,08	0	0	0,4	0	0,6	0,3	0	0,4	0,1	0	0,4	0	0,1	0,3	0,1	0,3	0,1	0,2	0,1	0,3	0,7	0,4	1	1	0,2	0,1
Cluster3	0,32	0	0	0	0,1	0	0	0,1	0	0,1	0	0	0,1	0,1	0	0,2	0	0	0	0,3	0,1	0	0,2	0	0	0,2	0,3
Cluster4	0,1	0	0	0	0,2	0	0,4	0	0	0	0,3	0,9	0,1	0	0,4	0,3	0	0	0,7	0	0,8	1,1	0,6	0,1	0	0,2	0,2
Cluster5	0,22	0	0,1	0	0	0	0,4	0,2	0,1	0	0	0	0,1	0,1	0,2	0,4	0,5	0,2	0,3	0,1	0,5	0,8	0,2	0	0,1	0,3	0
standard deviation		1,9	1,7	1,5	1,5	1,3	1,3	1,3	1,2	1,2	1,1	1,1	0,9	0,8	0,7	0,7	0,6	0,6	0,6	0,5	0,5	0,4	0,4	0,4	0,3	0,3	0,3

Figure 2: *Example of the proposed cluster based visualization in MS Excel*

cated by dark blocks of pixels along the diagonal. We have implemented this algorithm ourselves in R[8]. An example is displayed in figure 4. The distance measure we have used is Euclidean Distance. The intensity scale consisted of twelve shades of gray.

A dark cross appears in the top left corner of the ordered image. This corresponds to a part of the distance matrix containing zero values, which is of course the pairwise distance between two instances with zero values on all the features. A few of those instances exist, because of answers containing only function words (filtered out by the stop list) and very infrequent words (which are filtered out by the stringToWordVector filter in WEKA). Apart from these dark crosses, no dark blocks worth mentioning appear on the diagonal, which confirms that there is little or no cluster tendency in the data set.

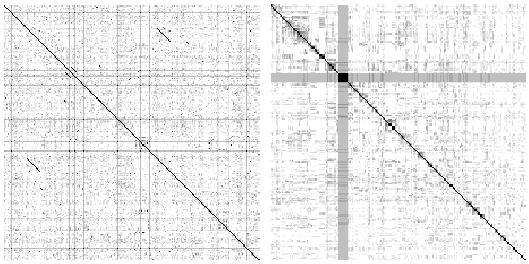


Figure 4: *Dissimilarity (left) and Ordered Dissimilarity (right) Image for Question 1*

6. Conclusions

We presented two experiments in attempting to detect arguments in replies to EU green papers. The first was aimed at classifying sentences as either argumentative or non-argumentative. From (Mochales and Moens, 2011) we learned that it should be feasible to automatically separate a text into argumentative and non-argumentative statements. Contrary to them we did not have a reasonably large tagged document set to train a machine learner. We resorted to a symbolic approach using keywords and regular expressions. Our classifier performs worse than theirs (F-score of 51 versus 73), probably partially due to difference in the type of documents. The Araucaria set that Mochales used is specifically aimed at argumentation and contains analysed arguments from newspapers, blogs and the like. Our set of replies to green papers is written in a far less argumentative style. Their second set consisted of documents extracted from legal texts of the European Court of Human Rights (ECHR), that has developed a standard type of reasoning and structure of argumentation over the years (Mochales and Moens, 2011). Our documents are written by different authors and their styles differ greatly.

In contrast to this first experiment, we found in our second series of experiments that semantic cohesion in the data is greater than cohesion based on linguistic constructs and vocabulary. This different result may have something to do with the different set of features used. Even though this result is promising, we must conclude that using content words in the answers to perform a clustering aiming at a semantic level of argument recognition was not feasible. This is partly due to the small size of the data set and the absence of a proper classification in the data. There appears to be no natural partitioning in the data, other than a very coarse topic-based division.

We are inclined to conclude that other features should be used to find any relevant grouping in this dataset. We will name a few possibilities here. Extending the work in our first experiment, the set of key words might be expanded with *argumentative phrases*, such as "First of all" or "as opposed to". Some research has been done on defining such phrases, see (van Eemeren et al., 2007) and (Knott and Dale, 1994). Some phrases may be grouped together, such as 'firstly' and 'secondly'. A related set of features could be created by tagging *sentiment phrases*, as has been described in (Fei et al., 2004).

One may also think of ways to tackle the problem of the small size of the data set. A model may be trained on an annotated argument corpus such as the Araucaria database. This would of course not take the specific terminology of a domain into account, but the model may be combined with a bag-of-words or an ontology to form a new model applying for both structural and symbolic classification. Furthermore, usage of the ART will lead to the creation of a corpus that can be used for future research.

To sum up, the results of our various experiments in automated support for finding and tagging arguments in natural language texts are not promising. The task seems too hard for the present state of the art, at least without a substantial corpus of tagged texts to use for training and testing.

The first step on this route therefore must be to set up such a corpus. The manual tagging of arguments using our tool is a logical step in that process. Making the ART available

[8] http://www.r-project.org/

as open source software, letting people tag arguments in responses to EU green papers and store these on our server will hopefully provide us with a usable corpus in the longer run. By making different levels of granularity available as described in section 2., the ART enables people to generate a gold standard at all these levels (that can be used as training and test set). This will enable experimentation with NLP techniques at any level. When automated support proves to be feasible, we can augment the existing user interface in such a way that users can benefit from it.

Acknowledgments
This research was partially funded by the EU in the Framework 7 programme (Grant Agreement No 247228) in the ICT for Governance and Policy Modeling theme (ICT-2009.7.3). Thanks to Sander Latour who helped with the first experiment described.

7. References

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, pages 1–33.

Bezdek, J. and Hathaway, R.). VAT: a tool for visual assessment of (cluster) tendency. *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, pages 2225–2230.

Brüninghaus, S. and Ashley, K. (2005). Generating Legal Arguments and Predictions from Case Texts. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law*, pages 65–74, New York, NY, USA. ACM Press.

Fei, Z., Liu, J., and Wu, G. (2004). Sentiment classification using phrase patterns. *The Fourth International Conference onComputer and Information Technology, 2004. CIT '04.*, pages 1147–1152.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, September.

Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse processes*, 18(1):35–62.

Marcu, D. and Echihabi, A. (2002). Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 368–375.

Maulik, U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, December.

Mochales, R. and Moens, M.-F. (2011). Argumentation mining. *Artif. Intell. Law*, 19:1–22, March.

Moens, M.-F., Boiy, E., Mochales, R., and Reed, C. (2007). Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 225–230, New York, NY, USA. ACM Press.

Moon, T. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA. ACM.

Slonim, N., Friedman, N., and Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02*, page 129, New York, New York, USA. ACM Press.

Strehl, A. and Ghosh, J. (2003). Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *The Journal of Machine Learning Research*, 3:583–617.

van Eemeren, F. H., Houtlosser, P., and Snoeck Henkemans, a. F. (2007). Indicators of argument schemes. In Eemeren, F. H., Houtlosser, P., and Henkemans, A. F. S., editors, *Argumentative Indicators in Discourse*, volume 12 of *Argumentation Library*, chapter 6, pages 137–191. Springer Netherlands, Dordrecht.

Walton, D. (2002). In *Legal Argumentation and Evidence*. Pennsylvania State University Press.

Network Analysis of Italian Constitutional Case Law

Tommaso Agnoloni
Institute of Legal Information Theory and Techniques ITTIG - CNR
Via de'Barucci 20, Firenze, Italy
agnoloni@ittig.cnr.it

Abstract

In this paper we report on an ongoing research on the application of network metrics to the corpus of Italian constitutional case law. The research was enabled by the recent release as open data of the complete datasets of the jurisprudential production of the Italian Constitutional Court. The datasets include the complete textual corpora of Court judgements together with a rich set of associated structured metadata. Using the standard unique identifiers for case law recommended by the EU Council and a recently developed jurisprudential reference extractor, adapted to constitutional case law texts, we were able to construct the graph of jurisprudential references of Italian constitutional decisions. On the resulting network, first metrics have been evaluated and further research activities are foreseen exploiting the richness of the datasets and their potential connections.

Keywords: network analysis, case law citations, open data

1. Introduction

In 2013 the Italian Constitutional Court released as open data the complete datasets of its decisions pronounced from its origin in 1956. According to the open data principles the data are released in open format (XML) and with an open license (CC-BY-SA-3.0). This is to the best of our knowledge the first massive release of reusable legal open data in Italy, opening unprecedented opportunities of datasets enrichment, interlinking and analysis under multiple point of views, with relevant interest among legal scholars and, for the relevance of the institution, to the wider public. In this paper we report on the activities of dataset analysis, enrichment and interlinking that allowed us to construct the Italian constitutional case law citation network. Preliminary results of network analysis on the resulting graph are described as well the planned activities for this ongoing research.

2. The Italian Constitutional Court case law in Open Data

The following datasets, complete and updated since the Court origin in 1956, have been released in the Open Data section of the constitutional Court website [1]:

- the archive of Court decisions
- the archive of the legal summaries (*massime*) edited by the Court itself resuming legal issues and motivations related to each decision
- biographic information on the constitutional judges composing the Court in its history
- registers about pending norms in front of the Court

The datasets are kept updated on a regular basis and their content is also accessible and searchable through the public web interface provided by the Court on its website.
Data are published in XML format structured according to the associated DTDs (also published) listing the available metadata fields and their relations.

[1] www.cortecostituzionale.it/ActionPagina_1177.do *retrieved Feb. 2014*

In the present work we restricted our analysis to the data published in the dataset of decisions and related *massime*. The dataset of decisions contains the full texts of judgements issued by the Court along with associated metadata:

- typology of decision (judgement, order or decree)
- year of publication
- number
- date of decision
- date of deposit
- names of judges composing the judicial panel
- name of the president judge

texts are structured in sections explicitly annotated in the XML document:

- heading of the document
- *fatto* : the facts originating the constitutional decision
- *diritto*: the legal issues raised by the case
- *dispositivo*: the decision given the facts and the legal issues

Each decision can have N legal summaries (*massima*), one for each legal issue dealt with in the decision. Each *massima* in the dataset has the following metadata associated to its text:

- metadata of the decision it refers to
- type of judgement
- subject (a list of concise titles, in free text, describing both the legal field and the outcomes of the decision)
- constitutional parameters (references to the norms object of the issue of constitutional legitimacy)

As a first dataset enrichment step we applied the standard unique identifier ECLI to each judgement. This is also the first step towards the evolution of the open dataset to the linked open data framework and its further interlinking.

3. ECLI and jurisprudential reference extraction

ECLI is the European recommendation of the European Council (EU-Council, 2011) establishing a standard identifier for case law (European Case Law Identifier). The identifier is composed of five fields in the following order: "ECLI" abbreviation, country code, court code, year of the decision, unique ordinal number of the decision, all separated by a colon (":"). The metadata associated with the decisions of the Court can be easily serialized to compose the ECLI in order to attribute to each decision its standard identifier. Given the authority abbreviation for the Constitutional Court (COST) and the codes describing the type of decision, used as prefix to the decision number (S for Judgement, O for order, D for decree), the ECLI is composed as follows

ECLI:IT:COST:{*year*}:{*decision_type*}{*number*}

3.1. Prudence

Following the introduction of ECLI, *Prudence*, a jurisprudential reference parser for Italian case law have been developed (Bacci et al., 2013) to extract jurisprudential references from plain text and serialize them in the ECLI standard format.

Originally developed for the extraction of jurisprudential references in civil case law of first instance, we tested and adapted *Prudence* on the texts of constitutional judgements to cover more lexical citation forms typically used in constitutional case law. In this preliminary investigation we were interested in the extraction of references to other constitutional judgements and discarded references to judgements of other (lower ranked) courts. Having the whole dataset at disposal and the exhaustive list of the ECLI identifiers of every decision of the constitutional court in history, we filtered the results of the automatic extraction by discarding malformed (not existing) extracted references and completing the identifiers of partially extracted references.

The evalution of the parser on a sample of 608 manually annotated citation contexts from a set of 60 cases evenly distributed over time resulted in a Precision of 98.4% and a Recall of 91.7% (see Table 1).

# documents	60
# citation contexts	608
# manually annotated references	1294
# correctly extracted references	1170
# wrong extracted references	18
# not extracted references	106
Accuracy	90.4%
Precision	98.4%
Recall	91.7%
F1	94.9%

Table 1: Evaluation of *Prudence* on the extraction of constitutional references

4. The Constitutional Court case law citation network

Several scholars have applied network analysis to case law citation networks both in common law (Fowler and Jeon, 2008) and civil law (Van Opijnen, 2012), (Winkels and de Ruyter, 2012) legal traditions. Our aim here was, as a first step, to provide the technical premises and an experimental testbed to test network metrics on Italian constitutional case law in order to allow further legal analysis and results comparison.

Thanks to the attribution of identifiers to decisions and to the extraction from text of jurisprudential references we were able to construct the overall citation network of the constitutional decisions of the Court. The nodes of the network are the ECLI of each decision; an edge among two nodes exists if a jurisprudential reference exists among the corresponding decisions. Edges are directed from the citing to the cited document and their weight is the number of references among the subtended nodes.

4.1. data analysis methodology

As a starting point the analysis was limited to a single type of judgement, selected by filtering the dataset by the corresponding attribute "*tipologia_giudizio*". Judgements "*in via incidentale*" are those originated by constitutional legitimacy exception on a norm, raised by a judge during a trial.

These are the majority of judgements of the Court. The other main typology of constitutional legitimacy exception "*in via principale*" is the one promoted by state institutions (government, parliament, regional government). The distribution of the typology of judgements is the one reported below:

Type of judgement	Percentage
in via incidentale	78.7%
in via principale	11.2%
other	10.1%
Total number of judgements	19085

Table 2: Distribution of judgement typology in the decisions dataset

See (Bellocci and Giovannetti, 2010) for more details.

On the selected judgements we started with the analysis of the network of internal references, *i.e.* references of the Court citing its precedents, and discarded for the moment jurisprudential references to other courts. The type of link (internal or external) is easily distinguished by the issuing authority in the ECLI (third field). The application of the jurisprudential reference parser on the selected corpus of judgements *in via incidentale* resulted in the distribution reported in Table 3

4.2. Preliminary results

From the collected data we were able to construct and visualize the Constitutional Court citation network using Gephi[2] (Fig 1).

[2] www.gephi.org

Type of reference	Percentage
constitutional	70.2%
to other courts	29.8%
Total number of references	139689

Table 3: Distribution of extracted jurisprudential references

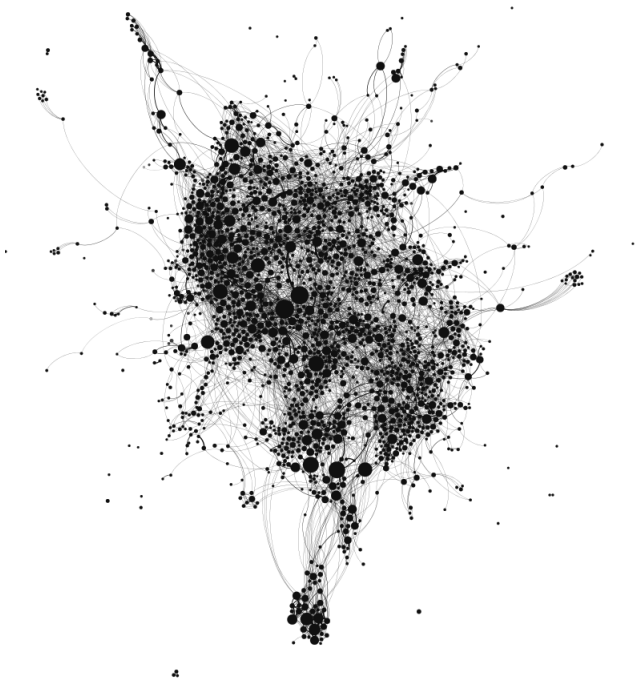


Figure 1: Citation network

The overall citation network consists of 98113 citations distributed among 14224 nodes. The basic topological properties of the resulting graph as computed with standard graph analysis algorithms are reported in Table 4.

Metrics	Value
Number of nodes	14224
Number of edges	38972
In-degree	[0-74]
Out-degree	[0-106]
Average degree	2.74
Network diameter	22
Average path length	6.96
Modularity	0.78
Number of communities	190

Table 4: Network properties

The first question we addressed is whether the distribution of the citation "edges" among the source nodes in the network adheres to the power law distribution predicted by the literature. This is to say that a small number of cases receive a large number of citations, and a large number of cases receive few citations or none at all. Not surprisingly for a man-made network like the web and many other studied citation networks where links are established by "preferential attachment" rather than by "random attachment", the distribution for our network actually exhibits a power law trend (Fig. 2).

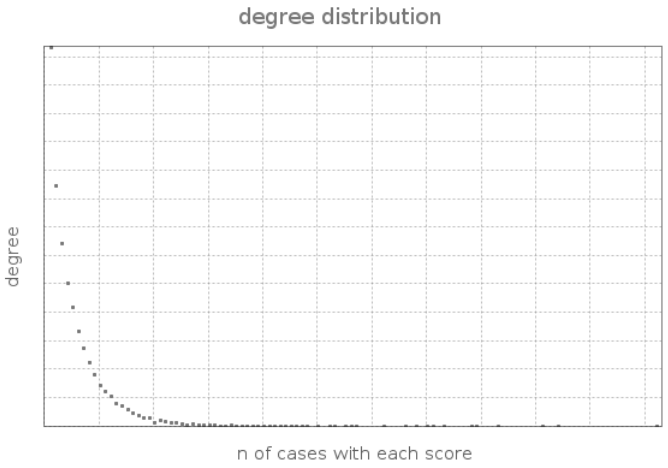


Figure 2: degree distribution

One of the goals of citation network analysis is to establish if and which metrics computed on the citation graph are indicator of the actual case law relevance or role of a case in the whole jurisprudential production of a court or, in general, in a given corpus. We started with the computation of two of the most significant, (Neale, 2013), (Van Opijnen, 2012) metrics for case law relevance: *degree* and *betweenness* centrality.

In a directed graph the *indegree* centrality measure simply counts for each node (judgement) the number of incoming edges (references). This represents a measure of absolute importance of a case. In Table 5 the top cited cases of the whole corpus are ranked according to their indegree.

Rank	Identifier	Value
1	ECLI:IT:COST:2004:S223	74
2	ECLI:IT:COST:2007:S26	62
3	ECLI:IT:COST:1990:S313	60
4	ECLI:IT:COST:1994:S240	52
5	ECLI:IT:COST:1995:S432	50
6	ECLI:IT:COST:1996:S131	49
7	ECLI:IT:COST:2007:S349	46
8	ECLI:IT:COST:1995:S313	44
9	ECLI:IT:COST:1996:S371	44
10	ECLI:IT:COST:1988:S971	43

Table 5: Top indegree

Betweenness centrality measures the number of shortest paths from all nodes to all others that run through a node. Intuitively nodes (documents) with a high betweenness centrality are those connecting different parts of the constitutional case-law graph. It is likely (to be

verified by legal analysis) that such decisions contain a court pronouncement over a general issue transversal to different subjects (e.g. procedural). Tab. 6 reports the top 10 decisions according to such measure.

Rank	Identifier	Value
1	ECLI:IT:COST:1996:S84	647238.85
2	ECLI:IT:COST:1995:S188	607207.46
3	ECLI:IT:COST:1980:O145	566919.13
4	ECLI:IT:COST:1995:S58	547260.24
5	ECLI:IT:COST:1993:S163	405385.8
6	ECLI:IT:COST:2007:S349	323632.52
7	ECLI:IT:COST:1995:S295	273148.31
8	ECLI:IT:COST:2007:S26	248356.67
9	ECLI:IT:COST:1993:S112	243448.15
10	ECLI:IT:COST:1994:S255	237633.67

Table 6: Top Betweenness Centrality

In order to prove their effectiveness as predictors of (legal) relevance, network metrics should be validated with respect to a benchmark obtained with different criteria. For example compared to the most searched cases in a legal database (Van Opijnen, 2012) or qualitatively matched with evidence provided by legal scholars based on their knowledge on the subject. This is one of the objectives of further development of this research along with the testing of other network measures.

Another important property of case law citation networks to be considered in further investigation is that they are dynamic networks, *i.e.* they vary over time. A more in depth analysis of the network should take into account the time element and include the analysis of the variation of the most important network features over time (Fowler and Jeon, 2008), (Neale, 2013), also correlated with known external events (e.g. legislative measures).

4.3. Analysis of subnetworks by topic

As seen in the dataset description of Sect. 2, each decision has *subject* metadata associated, reported in the field "title" of the *massime* dataset. Titles are manually attributed as free text to give an overview both of the content and the outcome of a decision. Overall an average of 6 titles are attributed to each massima. Titles can be a single keyword or a whole phrase. They require a lexical normalization at least to group different lexical forms for the same subject or concept. A work of further semantic clustering of related subjects is foreseen by exploiting automatic processing techniques. As an experiment we performed title normalization starting from the complete list of titles and manually grouped the reduced set around the broader topic of "*immigration/foreigner/refugee*". Based on grouped titles for this single case we were able to select the subnetwork by topic (Fig. 3). It is interesting to see overlapping cross citation among different topic subnetworks from in (black) and out (grey) nodes, and investigate the reasons from a legal point of view. This is also an intuitive way to visualize relevance and navigate case law precedents, also crossed with related contextual information (e.g. distribution per

Figure 3: subnetwork on the topic of "immigration"

year of cases on a topic, Fig. 4). This kind of analysis is for

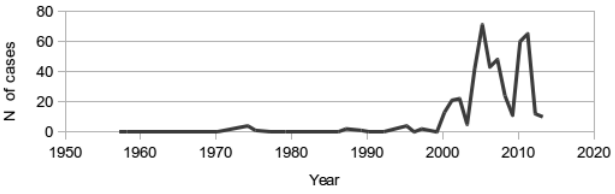


Figure 4: cases for the "immigration" topic per year

example interesting to give computable evidence to known historical and political facts (immigration have been actually a "hot" social and political topic in Italy from the end of the nineties).

5. Planned work

What we reported here are the results of very preliminary analysis made possible by the construction of the Italian constitutional case law citation network. Our plan is to go more in depth with the analysis under different perspectives:

data management

- expose the dataset enriched with the extracted features in a triple store or graph database in order to easily generate parametrized networks via queries (e.g. SPARQL queries) executed directly on the data (Hoekstra, 2013)

network analysis

- "slice" the graph according to the different facets available in the original dataset and analyse the evolution over time of both the structure of the network and of its main features (e.g. its hubs according to different centrality measures)

- use analysis of time-series network rankings for each case to determine "the age at which cases in the network typically cease to be important and what characteristics define those cases that continue to be important despite the passage of time" (Neale, 2013)

legal analysis

- check whether and how quantitative metrics on data provided by network analysis match existing constitutional law studies or suggest further legal considerations

linguistic and semantic analysis

- integrate the analysis of the *titles* in order to exploit the available subject annotation and identify network communities and subnetworks by legal topic, their properties and relations.

- derive a dataset associating outgoing references with their textual context in order to allow in depth linguistic analysis e.g. for the automatic recognition of the semantics of citations of precedents (in support, against, procedural) and eventually enable graph edges semantic labelling.

data mashup

- integrate the analysis with related datasets reported on a common timeline (e.g. interaction among constitutional decisions and the legislative process; links to national and regional legislation)

6. Conclusions

The application of quantitative methods in the legal domain requires wide availability of processable data to be applied. The open data release of the dataset of judgements of the Italian Constitutional Court gave us the opportunity to test network analysis and its metrics on a relevant branch of the Italian legal system. The network of citations have been constructed based on an enriched dataset where jurisprudential references among judgements have been automatically extracted from plain text by applying *Prudence* - a jurisprudential reference parser - to the whole corpus of decisions.

Further analysis, under different perspectives, of the graph obtained by interlinking the datasets released by the Constitutional Court is foreseen for the next phases of this research.

7. References

L. Bacci, E. Francesconi, and M.T. Sagri, 2013. *A Proposal for Introducing the ECLI Standard in the Italian Judicial Documentary System*, pages 49–58. IOS Press, Amsterdam (NL).

M. Bellocci and T. Giovannetti. 2010. Il quadro delle tipologie decisorie nelle pronunce della corte costituzionale. http://www.cortecostituzionale.it/studiRicerche.do.

EU-Council. 2011. Council conclusions inviting the introduction of the european case law identifier (ecli) and a minimum set of uniform metadata for case law. *Official Journal of The European Union*, 2011/C 127/01.

J.H. Fowler and S. Jeon. 2008. The authority of supreme court precedent. *Social networks*, 30(1):16–30.

R. Hoekstra. 2013. A network analysis of dutch regulations - using the metalex document server. In *Proceedings NAiL 2013 Workshop Network Analysisis in Law*, Held in conjunction with ICAIL 2013, Rome (Italy).

T. Neale. 2013. Citation analysis of canadian case law. *Journal of Open Access to Law (JOAL)*, 1(1).

M. Van Opijnen, 2012. *Citation Analysis and Beyond: In Search of Indicators Measuring Case Law Importance*, pages 95–104. IOS Press, Amsterdam (NL).

R. Winkels and J. de Ruyter. 2012. Survival of the fittest: Network analysis of dutch supreme court cases. In *Proceedings of the 25th IVR Congress Conference on AI Approaches to the Complexity of Legal Systems*, AICOL'11, pages 106–115, Berlin, Heidelberg. Springer-Verlag.