Automatic extraction of legal concepts and definitions

Winkels, R.G.F.; Hoekstra, R.J.

Published in:
Frontiers in Artificial Intelligence and Applications

DOI:
10.3233/978-1-61499-167-0-157

Automatic Extraction of Legal Concepts and Definitions

Radboud WINKELS [a,1], Rinke HOEKSTRA [a,b]
[a] *Leibniz Center for Law, University of Amsterdam*
[b] *Computer Science Department, VU University Amsterdam*

Abstract. In this paper we present the results of an experiment in automatic concept and definition extraction from written sources of law using relatively simple natural language and standard semantic web technology. The software was tested on six laws from the tax domain.

Keywords. NLP, Semantic Web, version management, Wordnet, MetaLex, annotation

1. Introduction

Public administrations realize that strong and explicit links between their data stores and business rules on the one hand and the sources of law that provide the basis for their existence on the other is essential. These links not only facilitate explaining and justifying their decisions and operations, it also improves impact assessment of legal changes and the ability to give feedback about problems in the implementation of these changes to policy makers and legislators.

The Dutch Tax and Customs Administration (DTCA) is responsible for implementing the tax- and customs legislation. Thus far, the administration that is conditional for this task is based more on usual conduct of business than on its basis in law. Careful analysis of legislation, looking for the concepts that are mentioned or defined and the relations between them, is a time and effort consuming task. Human analysts may make mistakes and two people may easily arrive at different conceptual models.

To what extent is automatic support feasible here? This question has been addressed in the past for classifying (normative) sentences or paragraphs in sources of law [1][2][5] and for suggesting model fragments for these sentences [6]. The first part is not sufficient for the purposes of the DTCA and the second part requires too much and heavy machine processing (using i.a. a full dependency parser for Dutch) and human intervention (i.a. to select the correct parse tree). In this paper we discuss the results of an experiment using simple natural language processing and standard Semantic Web technology to extract concepts, their relations and definitions from written sources of law in the tax domain. We will first specify the goals of the

[1] Corresponding author: Radboud Winkels, Leibniz Center for Law, University of Amsterdam, PO Box 1030, 1000 BA Amsterdam, Netherlands; Email: winkels@uva.nl

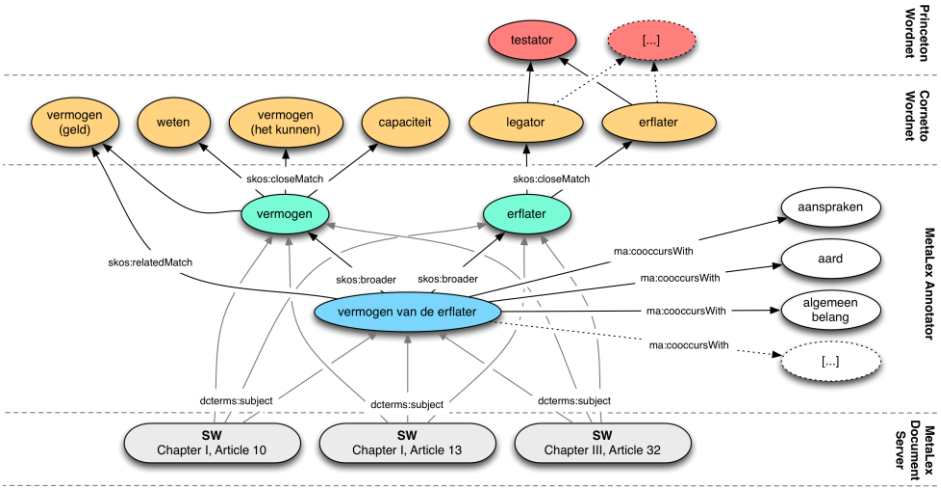

Figure 1: Some concepts and their relations found in the Succession Law (SW).

experiment, then the approach taken and the results obtained. We will end with conclusions and suggestions for improvements and future work.

2. Goals of the Experiment

The ultimate goal of the experiment is to create a conceptual model based on sources of law. A first step is to list all concepts as defined in legislation and possibly other sources. This list will form the basis for the implementation of administrative processes at the DTCA. Moreover, for every concept we need to add information:

1. A *reference* to the text where the concept is defined if that is the case in the input sources;
2. The *definition* of the concept if it is given in the sources;
3. References to all *occurrences* of the concept in the sources;
4. Possible *relations* with other concepts;
5. Possible *synonyms* of the concept or strongly related concepts.

To accomplish this task we designed and built the MetaLex Annotator (MA), a collection of Python scripts to parse legal documents and generate additional metadata. We will first describe some technical issues concerning MA and then discuss the extraction of concepts and definitions.

2.1. Use of Standards and Software

The MetaLex Annotator makes use of numerous standards. Sources of law are expected in a CEN MetaLex [2] compliant format. Recently we made all Dutch legislation

[2] CEN MetaLex is an open XML exchange format for legal and legislative resources, published as a CEN pre-norm. See: www.metalex.eu

available in that format at the MetaLex Document Server[3] [2]. The concepts that are found are represented as SKOS[4] concepts in RDF[5]. SKOS is a vocabulary that defines a number of basic classes and relations for expressing simple taxonomic information (such as broader and narrower relationships). We use the 'DC Terms' subset of the Dublin Core[6] standard for bibliographic annotations for linking concepts to the original sources. All concepts are stored in an RDF triple store, ClioPatria[7] and accessed using the standard RDF query language SPARQL[8].
The MetaLex Annotator is written in Python and uses an RDF library and the Natural Language Toolkit. For Part-of-speech tagging in Dutch we use the ConLL 2002 corpus 'ned.train'. We wrote our own simple grammar for Dutch, aimed at finding most noun phrases efficiently and unambiguously, because Dutch has a complex grammar that almost always leaves more than one interpretation of a sentence open. Our grammar is also conservative with respect to the length of noun phrases, to avoid too many false positives.

3. Concept Extraction

Concept extraction is simply implemented as follows (cf. [4]): For every article, parse every sentence individually:

- Every noun phrase refers to a concept;
- Every noun refers to a concept;
- Every noun within a noun phrase refers to a more general concept than the noun phrase does.

Take the following example from the Succession Law 1956:

Article 10, clause 9
The first clause is also applicable when a *debt* is part of the *estate of the testator* that came about as a consequence of a *testament*, to the extent that the *nominal value of that debt* is more than the *value* [...][9]

The concepts to be found are presented in *italic* and blue (light gray in black-and-white print). These are found using a regular expression grammar. The noun phrase 'estate of the testator' ("*vermogen van de erflater*" in Dutch) is a concept and the nouns 'estate' ("*vermogen*") and 'testator' ("*erflater*") are more general concepts; they are linked via the '`skos:broader`' relation (see Figure 1). These concepts are also linked to article 10 ("Artikel 10") with the '`dcterms:subject`' relation. What happens if the text of article 10 changes?

[3] http://doc.metalex.eu
[4] Simple Knowledge Organization System (SKOS), see: http://www.w3.org/2004/02/skos/
[5] Resource Description Framework (RDF), see http://www.w3.org/RDF/
[6] http://dublincore.org
[7] http://cliopatria.swi-prolog.org
[8] http://www.w3.org/standards/techs/sparql
[9] Original text in Dutch: "Het eerste lid is mede van toepassing, indien tot het *vermogen van de erflater* een *schuld* behoort, die is ontstaan als gevolg van een *uiterste wil*, voor zover de *nominale waarde van die schuld* meer bedraagt dan de *waarde* [...]"

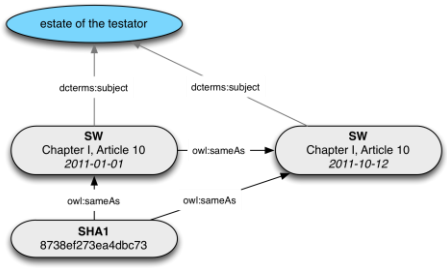

Figure 2: A hash code identifies the text of an article. A new version of the article appears at 2011-10-12 with the same hash code.

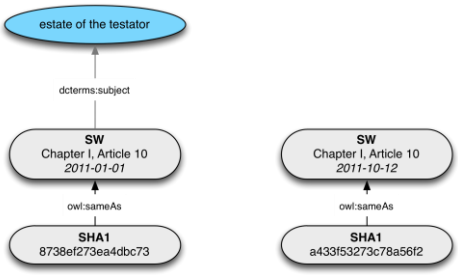

Figure 3: A new version of the article appears with a different hash code.

3.1. Version Management

The documents in the MetaLex Document Server come from the official Dutch portal of legislation: wetten.nl. The Basiswettenbestand (BWB) is the content management system for this portal. The current web service does not provide access to all versions of regulations (only to the latest), let alone at a level of granularity lower than entire regulations.[10] We therefore need some way of constructing a version history by regularly checking for new versions, and comparing them to those we looked at before. To uniquely identify the text of an article, a SHA1 hash[11] code is generated and attached to it. If the system receives a new version of the article with the same hash code, it assumes nothing has changed (see Figure 2). If the hash code differs, something definitely changed and the direct links with the earlier version and the concepts are broken (Figure 3). The different versions are still related at the 'work' level (a version independent identifier in CEN MetaLex).

3.2. Co-occurrence of Concepts

Concepts that occur together in an article are also registered and represented with a '`ma:cooccursWith`' relation, e.g. the concepts "*vermogen van de erflater*" (estate of the testator) and "*aanspraken*" (claims) in Figure 1.

3.3. Synonyms or Strongly Related Concepts

To find synonyms or strongly related concepts of the concepts that were found in sentences, we use Cornetto WordNet[12], an extensive thesaurus for the Dutch language. We use SPARQL queries to find matches between the (preferred) label of our concept and any in the Cornetto RDF store. Matches are linked to our concept using a '`skos:closeMatch`' relation (Figure 1). All composite terms are also linked to the

[10] Actually the situation is a bit more complicated than that. See [2] for details.

[11] SHA-1 is a cryptographic hash function designed by the US National Security Agency and published by the United States NIST as a U.S. Federal Information Processing Standard.

[12] http://www2.let.vu.nl/oz/cltl/cornetto/index.html

Cornetto concepts using a '`skos:relatedMatch`' relation (cf. "vermogen van de erflater" in Figure 1).

3.4. Results

We tested the MetaLex Annotator on 6 different sources of law in the tax domain, ranging from the 'income tax law' (more than 83,000 words) to the 'implementation decision of the Succession Law 1956'[13] (almost 2,000 words). The laws contain 1133 articles in total. It found 6,875 different concepts, occurring 22,681 times in total, 13% of the number of words in a text. Note that some words are part of several concepts since the constituting parts of longer concepts are concepts themselves (see above). Table 1 summarizes the results.

MA finds many concepts in a source of law, too many sometimes. Most of these time it concerns 'too long' concepts despite the conservative grammar, e.g. 'concepts' of 21 words like "*kader van een regeling voor onderling overleg op grond van het verdrag ter afschaffing van dubbele belasting in geval van winstcorrecties*" ('scope of a regulation for coordinated interventions acting on the treaty for abolishment of double taxing in case of profit adjustment'). It is certainly a noun phrase and it may even be considered a concept, but whether it is a very useful concept for the DTCA is questionable.

Long concepts are also created because of disjunctions; two or more concepts are also seen as one (the constituting parts are represented as separate concepts as well), e.g. concepts like "*behalen van belastbare winst uit onderneming <u>of</u> belastbaar resultaat uit overige werkzaamheden*" ('gaining of a taxable profit from enterprise <u>or</u> taxable result from other business').

Table 1: Number of words and found occurrences of concepts per law

Source	Nr of Words	Nr of Concepts	Concepts/ Words
Income tax law 2001 (BWBR0011353)	83,796	10,887	13%
General administrative law (BWBR0005537)	39,329	5,940	15%
General law concerning central government tax (BWBR0002320)	24,120	3,144	13%
Succession law 1956 (BWBR0002226)	13,980	1,898	14%
Implementation regulation donation- and inheritance law (BWBR0027018)	3,462	507	15%
Implementation decision of the Succession Law 1956 (BWBR0002227)	1,984	305	15%
Total	**166,671**	**22,681**	**13%**

Long concepts may not always be interesting, very short ones are neither. Table 2 presents the ten most frequent concepts in these six sources. These are general 'legal' or 'legislative' terms like 'application' and 'decision' and only 'amount' and

[13] "Uitvoeringsbesluit Successiewet 1956" in Dutch.

'(calendar) year' would be considered slightly specific for the tax domain. That most of the terms are indeed not very specific can also be seen from the IDF[14] scores.

Table 2: Absolute frequencies, number of documents and IDF scores of top-10 terms

	Term Dutch	English translation	Freq	Nr Docs	IDF
1	toepassing	application	1091	523	0.77
2	bedrag(en)	amount(s)	587	250	1.51
3	jaar/jaren	year(s)	456	210	1.69
4	betrekking	concerns	428	233	1.58
5	beschikking(en)[15]	decision(s)	382	173	1.88
6	besluit(en)	decision(s)	362	170	1.90
7	kalenderjaar/jaren	calendar year(s)	350	134	2.13
8	beroep	appeal	346	142	2.08
9	artikel(en)	article(s)	318	165	1.93
10	termijn(en)	period(s)	300	159	1.96

To see whether a term is specific for a document in a collection of documents, one typically uses the TF-IDF[16] score. The occurrences of 'application' in Table 2 in individual documents (articles) have TF-IDF scores between 0.05 and 0.77, so very low indeed. Terms with the highest TF-IDF scores are those that occur only in one article in the collection, like 'home help' ("*gezinshulp*") in art. 6.17 of the Income Tax Law (TF-IDF of 7.03). There are many of these; a list of highest TF-IDF score terms would be too long for this paper.

Table 3: Most frequent terms in one document

	Term Dutch	Source	English translation	TC	TF-IDF	Nr Docs
1	nederland	Income Tax Law art. 7.2	netherlands	37	2.57	87
2	woning	Income Tax Law art. 3.111	domicile	34	2.61	83
3	woning	Income Tax Law art. 3.119a	domicile	31	2.61	83
4	aandelen	Impl. Succession Law art. 8	stocks	25	2.84	66
5	auto	Income Tax Law art. 3.20	car	24	5.24	6
6	schulden	Income Tax Law art. 3.120	debts	24	3.51	34

[14] IDF – Inverse Document Frequency, a measure of whether the term is common or rare across all documents, calculated as the *log* of the total number of documents divided by the number of documents containing the term. The higher IDF, the more specific the term.

[15] A few of the 'beschikking' actually refer to 'placing at someone's disposal' ("*ter beschikking*" in Dutch).

[16] "Term Frequency-Inverse Document Frequency", it compensates for document length and overall frequencies of words in the collection. The higher the TF-IDF, the more specific the term is for the particular document.

7	jaar	Decision Suc. Law art. 5	year	24	2.11	137
8	zelfstandigen-aftrek	Income Tax Law art. 3.76	independents deduction	23	5.65	4
9	partner	Income Tax Law art. 2.17	partner	23	2.77	71
10	gebouw	Income Tax Law art. 3.20a	building	22	5.42	5

Table 3 shows the terms that occur the most in one document (article), e.g. "*nederland*" (Netherlands) is mentioned 37 times in article 7.2 of the Income Tax Law and has a TF-IDF score of 2.57. The term appears in 87 documents of our collection.

Nr laws	Freq. of co-occurrences
1	389,218
2	20,652
3	3,680
4	637
5	131
6	21
Total	414,339

When we look at the co-occurrence of concepts, we see similar results. For the six laws in our test set, we find 414,339 co-occurrences; 21 of these appear in all six laws and are not very specific. Examples are 'amount' ('*bedrag'*) - 'application' ('*toepassing'*) and 'request' ('*verzoek*') – 'decision' (*'beschikking'*). As one would expect, co-occurrences that appear in less laws are more interesting, like 'personal data' ('*persoonsgegevens'*) – 'basic administration' ('*basisadministratie'*). The ranking of frequencies of co-occurrences on both the level of laws (Table to the left) as on the level of articles follows a Zipf-like distribution.[17]

4. Definition Extraction

In earlier work by de Maat [5], definitions in Dutch legislation were found by looking for typical patterns in the text like "is understood by". He reports using 5 of 14 definition patterns to correctly classify sentences as definitions (92% recall, [7]). Interestingly, a SVM only scores a recall of 57% on definitions (95% recall on the total classification task – definitions only make up about 2% of the corpus). De Maat distinguishes so called 'type extensions' from definitions. They are very similar, but instead of completely defining a new term, they expand or limit an earlier definition (using words like "also" and "not"). When suggesting model fragments for classified definitions, it consists of three parts: the *definiendum* and the *definiens*, and, optionally, a scope declaration stating for which sources of law the definition applies. Most often, the scope is the particular law it is in ("this law") or "this law and the stipulations based on it".

We take a similar approach as can be seen from the example below. The extracted information is presented in Table 4. It shows the 'concept' (definiendum) and the 'definition' (definiens), and optional elements: 'modifier' to deal with type extensions, 'scope' for scope declarations and 'condition' for potential conditions for the definition to apply.

[17] Zipf's original law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.

Succession Law 1956, Article 1, clause 2
For the application of this law it is also understood by acquisition by inheritance law, the acquisition of licences and claims at or after the death of the testator if that acquisition is directly connected to the circumstance that the testator possessed these licences and claims.[18]

Table 4: Example extracted definition

Concept	acquisition by inheritance law
Definition	the acquisition of licences and claims at or after the death of the testator
Modifier	also
Scope	for the application of this law
Condition	that acquisition is directly connected to the circumstance that the testator possessed these licences and claims

4.1. Results

We tested the definition extraction on the same 6 laws and results are somewhat disappointing. MA finds definitions with few false positives (5%), but recall is only 42%. This is partly due to missing patterns for legal fictions or deeming provisions, e.g. the pattern '*wordt geacht*' ('is considered'). The main problem however is definitions in lists, which the patterns cannot handle, e.g.:

Succession Law 1956, Article 35c[19]
1. For the application of this chapter and the stipulations based upon it, it is understood by acquisition of enterprise wealth the acquisition of:
    a. an enterprise as meant in article 3.2 of […]
    b. a joined right as meant in […]
    c. wealth constituents […]
etc.

5. Conclusions and Discussion

The automatic recognition of concepts in legislative texts is feasible, but not perfect (yet). Dutch is a difficult language and far less parsing tools are available than for e.g. English. The relative simple and standard software we use enables us to find a lot of concepts, including a number of *false positives*, noun phrases that are incorrectly identified as concepts. Most of the time it concerns 'too long' concepts despite the conservative grammar. The stemming of words can be improved as well.

[18] Original text in Dutch: "Onder verkrijging krachtens erfrecht wordt voor de toepassing van deze wet mede verstaan de verkrijging van vergunningen en aanspraken bij of na het overlijden van de erflater indien die verkrijging rechtstreeks verband houdt met de omstandigheid dat de erflater die of dergelijke vergunningen en aanspraken bezat."

[19] Original text in Dutch: "Voor de toepassing van dit hoofdstuk en de daarop berustende bepalingen wordt onder de verkrijging van ondernemingsvermogen verstaan de verkrijging van:"

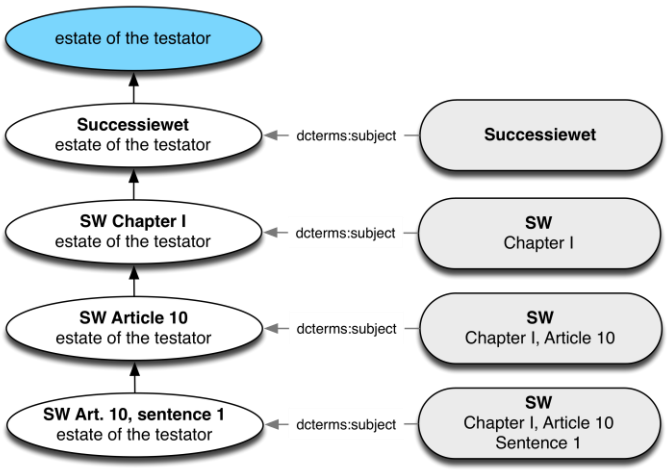

Figure 4: Concepts linked at different levels to source text

Both the very long and the very short concepts appear not to be very interesting for the people working in the tax administration. We probably should focus our report on the concepts 'in the middle'. It remains to be seen whether we should choose these based on the term count in documents (articles) within a law or TF-IDF scores and whether it is better to use a mean or median score (cf. Figure 5). We will evaluate this with users in the tax administration.

The linking to related terms based on Cornetto Wordnet works fine but is currently very slow. It also needs to be evaluated at which level the concepts are best linked to the original sources. Currently we do so at the article level, but it could also be done at clause or even sentence level. This is related to the *scope* of concepts (cf. the scope of definitions): If both article 10 and article 32 of a law use term X in their text, can one safely assume these terms refer to the same concept? Is the concept 'estate of the testator' ("*vermogen van de erflater*") in articles 10, 13 and 32 of the Succession Law the same concept (Figure 1)? A representational solution for this problem is the creation of a unique concept-id for every occurrence in the legislative text (Figure 4). One has to decide at which level occurrences of a concept need to be distinguished.

The extraction of definitions is less successful than the results reported by de Maat [6], even though we use the same method and patterns. Apparently the laws in the tax domain make more use of lists in definitions than laws from other domains and it uses some new patterns.

Acknowledgements
We would like to thank the Dutch Tax and Customs Administration for supporting part of this research.

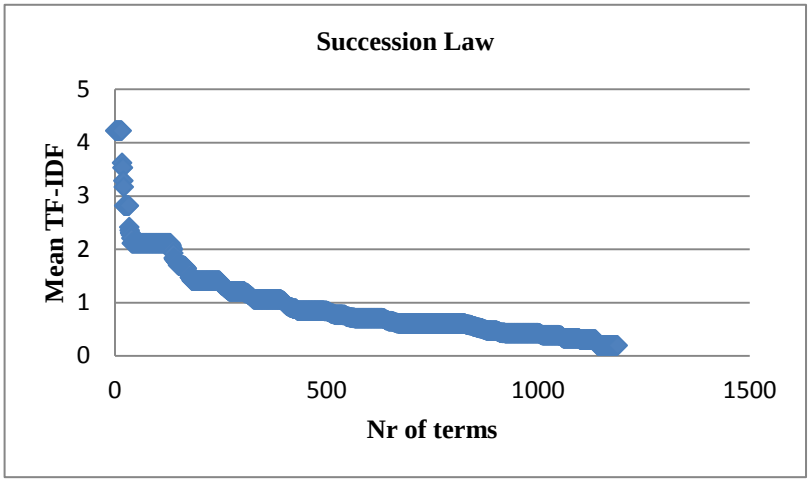

Figure 5: Frequency distribution of terms with mean TF-IDF scores for the Dutch Succession Law

References

[1] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni and C. Soria. Automatic semantics extraction in law documents. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL '05)*, pp. 133-140, ACM Press, New York, 2005.

[2] Gonçalves, T. and P. Quaresma. Is linguistic information relevant for the classification of legal texts? In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law (ICAIL '05),* pp. 168-176, ACM Press, New York, 2005.

[3] Hoekstra, R. The MetaLex Document Server - Legal Documents as Versioned Linked Data, *Proceedings of the International Semantic Web Conference (ISWC2011)*, pp. 128-143. Springer, Berlin, 2011.

[4] Jiang, X. and A-H Tan. Mining Ontological Knowledge from Domain-Specific Text Documents. In Proceedings of the Fifth IEEE International Conference on Data Mining, IEEE Computer Society, pp. 665-669, 2005.

[5] Maat, E. de, Winkels, R. Automatic Classification of Sentences in Dutch Laws. In Francesconi, E., Sartor, G., Tiscornia, D. (eds.) *Legal Knowledge and Information Systems. Jurix 2008: The Twenty-First Annual Conference*, pp. 207-216, IOS Press, Amsterdam, 2008.

[6] Maat, E. de, R. Winkels, and T. van Engers,. Making Sense of Legal Texts. In G. Grewendorf & M. Rathert (eds), *Formal Linguistics and Law*. Mouton deGruyter, Berlin, pp. 225-255. Series Trends in Linguistics. Studies and Monographs (TiLSM), 2009.

[7] Maat, E. de. *Making Sense of Legal Texts*. PhD thesis, University of Amsterdam, Amsterdam, Netherlands, 2012.