Domain Adaptation of General Natural Language Processing Tools for a Patent Claim Visualization System

Linda Andersson, Mihai Lupu, and Allan Hanbury

{andersson,lupu,hanbury}@ifs.tuwien.ac.at

Abstract. In this study we present a first step towards domain adaptation of Natural Language Processing (NLP) tools, which we use in a pipeline for a system to create a dependency claim graph (DCG). Our system takes advantage of patterns occurring in the patent domain notably of the characteristic of patent claims of containing technical terminology combined with legal rhetorical structure. Such patterns make the sentences generally difficult to understand for people, but can be leveraged by our system to assist the cognitive process of understanding the innovation described in the claim. We present this set of patterns, together with an extensive evaluation showing that the results are, even for this relatively difficult genre, at least 90% correct, as identified by both expert and non-expert users. The assessment of each generated DCG is based upon completeness, connection and a set of pre-defined relations.

Keywords: Graph visualization, domain adaptation, Natural Language Processing.

1 Introduction

The overall aim of this novel Dependency Claim Graphs (DCG) system is to support the cognitive process of reading and interpreting the claim text of a patent document. A patent document consists of four main textual components (title, abstract, description, and claim), intended to fulfil different communication goals. The claim has its own very special conceptual, syntactic and stylistic/rhetorical structure. It needs to be composed in such a way as to completely describe the essential component of the invention, while making patent infringement difficult [1].

In the IR community, the research focus has mainly been on improving and developing methods and systems for supporting patent experts in the process of Prior Art search (i.e. retrieving patent documents which could invalidate the patent) [2]. There have been two main evaluation campaigns (NTCIR and CLEF-IP) with patent-related tasks, while others (e.g. TREC-CHEM) have used substantial patent collections [3]. Less research attention has been given to other information processing activities conducted by the professional patent searchers. In order to formulate complex search queries, the patent experts extract phrases and terminologies used in the patent application. Part of this pre-search analysis consists of examining the claim section, to define scope and limitation [2, 4].

Additionally, during a post process analysis, patent claims are important in order to establish similarity between different patents. This motivates our efforts in developing

M. Lupu, E. Kanoulas, and F. Loizides (Eds.): IRFC 2013, LNCS 8201, pp. 70–82, 2013.

a system that supports the information analysis process by visualizing differences and similarities within a patent claim in order to show different aspects of the invention.

In order to generate claim graphs for the entire claim as well as for each paragraph, we use several different layers of linguistic information: Part-of-speech (PoS-tagger), phrase boundaries (chunker) and discourse theory. Instead of using a full-scale parser assigning syntactic relation, we use generic lexico-syntactic patterns for Relation Extraction (RE). We also use lexico-syntactic patterns to adapt the analysis from the NLP tools used in the pipeline to better reflect the syntax of claims sentence.

There are two main reasons for using lexico-syntactic patterns instead of a full-scale parser: first, the lack of robustness of the parser tagger and chunker, and second, the speed. Moreover, if we just change the focus from the mainstream genre such as newspaper articles to more specific corpora, several of the existing tools show a significant decrease in performance [5]. In a German study the accuracy decreased approximately by 5 percentage units (97% to 92%) when training on ideal German corpora and then testing on German Web corpora [6].

There are few NLP tools adapted for the patent domain. Furthermore, such tools are generally restricted to only working on pre-defined technical fields, as in [7] or [8]. Since the aim with the DCG system is to support the cognitive process of reading claims, the NLP applications used in the pipeline are required to handle all types of claims, as well as all technical fields. Therefore we investigate the use of generic lexico-syntactic patterns.

We used the English part of the CLEF–IP 2012 Passage Retrieval topic set as test collection, since the technical field distribution reflects the collection composition.

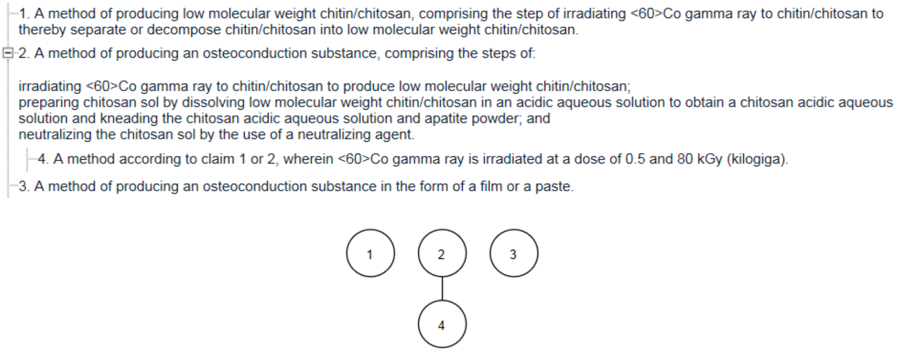
The rest of the paper is organized as follows. Section 2 gives insight to the related work and linguistic characteristics of the patent genre. Our method, the experiment and evaluation schema are presented in Section 3. In Section 4 the outcome of the evaluation task is presented along with analysis of general errors made by the NLP tools used in the pipeline. Conclusion and final remarks are given in Section 5.

2 Related Work

In order to create support tools for the patent expert we need to understand their daily work task, as well as the patent lifecycle and the linguistic character of this text genre [4].

The patent claims define the technical boundaries that should be protected by the patent. Therefore, the claims vocabulary consists of terms with legal impact as well as technical terms. The rhetorical structure of a claim is pre-defined into three parts: preamble, transitional phrase, and body. Ferraro [9] gives a more linguistic description of each part's function compared to the patent regulation literature. Here, we choose to use Ferraro's definition. The preamble is the claims introduction clause, which could include the main function of the invention as well as its purpose and field. The transitional phrase (or linking words), connect the preamble to the part specifying the invention itself (the body). In the transitional phrase words such "comprising", "containing", "including", "consisting of", "wherein" and "characterized in that" are frequently used. The body explains the invention and enumerates the legal and technical limitations.

Claims can be divided in two different categories: independent and dependent claims. An independent claim is a legal statement of its own, and does not refer back explicitly or implicitly to any other claims. A dependent claim depends on the claim/claims it explicitly refers back to by phrases such as "according to claim 1", "according to any of the previous claims", etc. Figure 1 displays the entire claim section and a claim tree according to the European Patent Office (EPO) existing tree claim structure of the patent application EP1306390 (A1).


1. A method of producing low molecular weight chitin/chitosan, comprising the step of irradiating <60>Co gamma ray to chitin/chitosan to thereby separate or decompose chitin/chitosan into low molecular weight chitin/chitosan.

2. A method of producing an osteoconduction substance, comprising the steps of:

irradiating <60>Co gamma ray to chitin/chitosan to produce low molecular weight chitin/chitosan;
preparing chitosan sol by dissolving low molecular weight chitin/chitosan in an acidic aqueous solution to obtain a chitosan acidic aqueous solution and kneading the chitosan acidic aqueous solution and apatite powder; and
neutralizing the chitosan sol by the use of a neutralizing agent.

4. A method according to claim 1 or 2, wherein <60>Co gamma ray is irradiated at a dose of 0.5 and 80 kGy (kilogiga).

3. A method of producing an osteoconduction substance in the form of a film or a paste.

Fig. 1. Claim tree of Patent Application EP-1306390 (A1)[1]

The application EP1306390 (A1) consists of three independent claims (1, 2, 3) and one dependent claim (4), as visualized in Figure 1. In comparison with the EPO claim tree which make use of explicit reference in the text to other claims i.e. "according to previous claims", "according to claim 2 and 1" etc. our DCG system takes advantage of implicit reference by detecting discourse references (e.g "the chitosan acidic aqueous solution" is referring to the same entity as "a chitosan acidic aqueous solution"). However, in order to detect noun phrases (NP) we first need to parse each sentence in order to identify given entities. Figure 2 shows the sentence claim graph of claim 1.

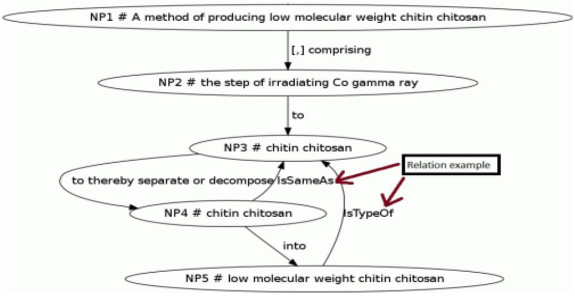

NP1 # A method of producing low molecular weight chitin chitosan
[,] comprising
NP2 # the step of irradiating Co gamma ray
to
NP3 # chitin chitosan
Relation example
to thereby separate or decompose
IsSameAs
NP4 # chitin chitosan
IsTypeOf
into
NP5 # low molecular weight chitin chitosan

Fig. 2. Claim 1 of Patent Application EP-1306390 (A1)

[1] http://bit.ly/19z9t7V

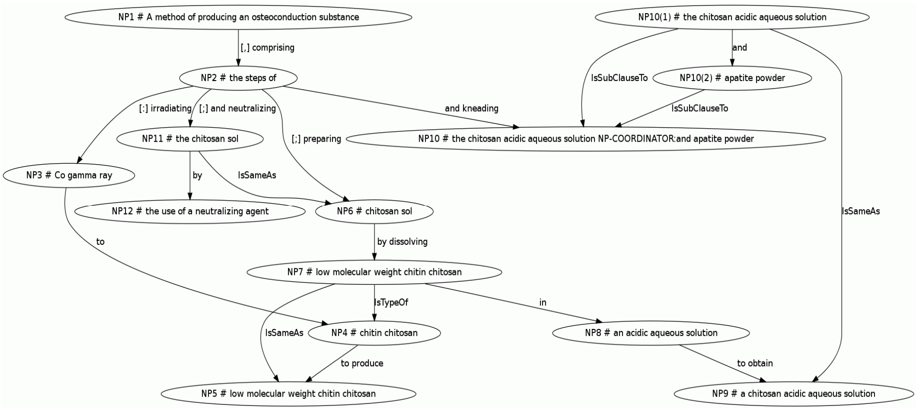

NP1 # A method of producing an osteoconduction substance
[,] comprising
NP2 # the steps of
[:] irradiating
[;] and neutralizing
[;] preparing
and kneading
NP10(1) # the chitosan acidic aqueous solution
and
IsSubClauseTo
NP10(2) # apatite powder
IsSubClauseTo
NP10 # the chitosan acidic aqueous solution NP-COORDINATOR:and apatite powder
NP11 # the chitosan sol
NP3 # Co gamma ray
by
IsSameAs
NP12 # the use of a neutralizing agent
NP6 # chitosan sol
IsSameAs
to
by dissolving
NP7 # low molecular weight chitin chitosan
IsTypeOf
in
IsSameAs
NP4 # chitin chitosan
NP8 # an acidic aqueous solution
to produce
to obtain
NP5 # low molecular weight chitin chitosan
NP9 # a chitosan acidic aqueous solution

Fig. 3. Claim 2 of Patent Application EP-1306390 (A1)

The aim of the graph for each claim sentence is to identify the boundary of the main NP, as well as assign discourse relations such as IsSameAs and IsTypeOf. Each node is assigned a sequence marker NP1, NP2 etc. The more complex claim sentence (i.e. Claim 2) is presented in Figure 3.

In Figure 3 we can also see a larger NP constituent, consisting of a set of smaller NPs combined with a coordinator and collapsed into node NP10. The smaller NPs (Sub nodes) are linked by an IsSubClauseTo relation to the main node. The linear order of the sentence is changed since the sentence has a transitional phrase that indicates a listing of activities.

In the full scale DCG system, each claim sentence tree will be interactively connected to other claims when the user requests links based upon different pre-defined relations such as explicit dependency relations (i.e. IsDependencyOf), IsSameAs and IsTypeOf. However, in this present paper we evaluate the effectiveness of the pre-process i.e. the construction of the tree per claim sentence.

2.1 Characteristics of the Patent Claim Text

Verberne et al [10] presented a comparative linguistic genre study comparing patent claims text and sentences with general language resources. They found that the sentences of patent claims (allowing semicolon and colon as sentence splitter) were generally longer than the sentences found in the British National Corpus (BNC). The patent sentences had a median of 22 tokens and an average length of 53 tokens (based upon a comparative study consisting of 581k sentences). Ferraro [7] reported it is not unusual to have sentences in the claim section consisting of 250 tokens; and in Wäschle and Riezler [11] it was also reported that a corpus of 500k claim sentences consisted of approximately 18,355,584 tokens and 270,013 types and the average type-token ratio (TTR) is 0.0147. The TTR indicates the variation in vocabulary distribution, which needs to be handled in the NLP pipeline.

A few domain adaptations of NLP tools have relied on incorporating domain knowledge in the NLP process by extracting terms from patent collections [1, 12]. However, to just increase the lexical coverage will not solve the problem, since token coverage is only part of the problem. Verberne et al [10] concluded there were no significant differences between general English and the English used in patent claim text for single token coverage, the (new) technical terminology is more likely introduced on the multi-word level consisting of complex NPs. Also, the literature addressing terminology extraction confirms that the majority of the technical dictionaries consist of terms with more than one word [13]. The technical multi-word phrases consist of noun phrases containing common adjectives, nouns and occasionally prepositions (e.g. 'of'). Therefore, it is important that the focus for domain adaptation lies within identifying correct noun phrase boundaries as well as increasing of lexical coverage.

2.2 Domain Adaptation and Evaluation

In [1], the aim was to reduce complexity in claims and increase readability. It required modification of the pre-processing, training of a super tagger (lexical driven) and additional domain rules to define dependency relations. Previous work addressing claim readability has been conducted on Japanese patent claims [14].

The retrieval system PHASAR has been domain adapted towards the patent domain by increase of lexicon coverage [12]. The system integrates linguistic notation in the search mechanism, it uses linguistic information and displays linguistic knowledge to the searcher. The system aims to capture dependency relationships between words via dependency triples. PHASAR uses a special grammar based on AEGIR (an extension of Context Free grammar formalism) adapted for robust parsing to be used in IR.

In terms of evaluation, to the best of our knowledge, only small-scale linguistic evaluation of parsers has been conducted. In [10], a comparison between two different parsers was performed: AEGIR used in the PHASER system, and the Connexor CFG parser. Conducting a linguistic evaluation of the performance of a parser or part of speech application is a time consuming task and requires both linguistic expertise as well as domain knowledge. In [10], 100 randomly selected short patent sentences (5-9 words) were assessed based upon generated dependency triples; the F1-scores for AEGIR 0.47 and for Connexor CFG 0.71 were calculated. The inter-annotator agreement was 0.83 and was computed by counting the number of identical triples divided by the total number of triples created by another annotator.

In another study, Parapatics and Dittenbach also aimed to reduce the complexity in claims [8], the General Architecture for Text Engineering (GATE) was used to decompose sentences by identifying the claim-subject and assigning dependent claims to the correct independent claim. The Stanford dependency parser was used in the large evaluation (5000 claim sentences), but only its performance in terms of ability to parse and its memory usage was assessed for the decomposed and for the original sentences. The correctness of the parsed sentences was never investigated.

In Ferraro [7], Minipar was used as part of the NLP-pipeline to extract verbal relations in patent claims. No evaluation of the parser performance on the patent text was conducted, only citing the performance of Minipar on the Susanne Corpus[2] (0.89 precision). However, before the final parsing of the patent sentence, the sentence was decomposed to smaller units by using a domain adapted segmentation tool, as well as a rules driven algorithm for paraphrasing the segment into complete sentences [15].

3 Our Approach

The English part of the CLEF–IP 2012 Passage Retrieval topic set was used as training and test set. For training purposes we randomly selected 40 claim sentences, which we manually investigated. The test set consisted of 600 randomly selected claim sentences. In Table 1 the average number of tokens and types, as well as the average TTR is shown for the test set, divided per International Patent Classification system (IPC) section. In parenthesis, for each average, we indicate the standard deviation, to give an idea of the range in the entire population. We follow the same convention throughout this paper.

Table 1. Token, type TTR distribution for Test Set

IPC	Token Average (stddev)	Type Average (stddev)	TTR Average (stddev)
A (Human Necessities)	33.37 (17.10)	26.34 (9.55)	0.84 (0.13)
B (Performing Operation; transporting)	33.59 (20.96)	26.37 (11.75)	0.85 (0.13)
C (Chemistry; Metallurgy)	28.60 (14.91)	23.95 (9.36)	0.88 (0.11)
D (Textiles; Paper)	41.53 (20.29)	31.81 (11.76)	0.81 (0.11)
F (Mechanical Engineering; Lighting; Heating,Weapons; Blasting)	38.75 (25.00)	29.29 (13.62)	0.81 (0.12)
G (Physics)	30.26 (19.25)	23.99 (10.58)	0.86 (0.13)
H (Electricity)	38.16 (23.33)	27.89 (12.21)	0.81 (0.85)
Total	33.21 (19.31)	26.16 (10.88)	0.85 (0.13)

In a previous work addressing phrase retrieval [16], an observation study of noun phrase patterns was conducted in order to define noun phrase boundaries in the patent domain.

However, the DCG system requires more flexibility in the phrase boundary than given by the lexico-syntactic pattern. Therefore we chose to use the baseNP Chunker [17]; and construct generic rules modifying the output of the chunker based upon previously observed patterns. All sentences were annotated with the Stanford Part-of-Speech tagger, using the english-left3words-distsim.tagger model [18].

[2] http://www.grsampson.net/SueDoc.html

Table 2. Assessment of evaluation parameters by all assessors

Rule	Original NP Sequence	Modified NP Sequence	Modifying
"said" as an article	said/VBD [supercritical/JJ fluid/NN]	[said/VBD supercritical/JJ fluid/NN].	PoS-tagger
preposition within the preamble phrase	[The/DT soccer/NN shoe/NN] of/IN [claim/NN 4/CD]	[The/DT soccer/NN shoe/NN of/IN claim/NN 4/CD]	Chunker
include present participle	[A/DT method/NN] of/IN fabricating/VBG [a/DT semi-conductor/NN device/NN]	[A/DT method/NN of/IN fabricating/VBG a/DT semi-conductor/NN device/NN]	Chunker
infinitive verb tagged as NN	[said/VBD laser/NN radiation/NN] to/TO [exit/NN] [said/VBD exit/NN system/NN]	[said/VBD laser/NN radiation/NN] to/TO exit/VB [said/VBD exit/NN system/NN].	PoS-tagger
include digits into the NP	NP [The/DT method/NN of/IN any/DT of/IN claims/NNS] [12/CD to/TO 16/CD]	[The/DT method/NN of/IN any/DT of/IN claims/NNS 12/CD to/TO 16/CD]	PoS-tagger
list of NPs	in [the/DT group/NN] consisting/VBG of/IN [a/DT photoresist/NN] ,/, [a/DT photoresist/NN residue/NN] ,/, and/CC [a/DT combination/NN]	into [the/DT group/NN] consisting/VBG of/IN [a/DT photoresist/NN ,/, a/DT photoresist/NN residue/NN ,/, and/CC a/DT combination/NN]	Claims discourse adaptation specific rules
	A sub rule to 7, Identifying, transition phrases listing sub clauses as seen in figure 2		

We created 9 main rules to adjust and adopt the output from the PoS-tagger and chunker to better reflect the patent domain (see Table 2).

These rules were implemented in order to generate a connected graph and to better identify the NP boundaries occurring in patent claims. Before submitting the claims text to the PoS-tagger and baseNP Chunker, a generic abbreviation handler and a modified sentence splitter were applied. Also, special signs were removed.

3.1 Nodes and Relations

The graph nodes reflect the sentence noun phrases, where we have chosen to collapse NPs consisting of smaller NPs into larger complex NPs, as seen in Figures 2 and 3. The relations (or links) consist mostly of the sentence's verb, preposition or other transition function such as clause markers (‘,’, “;”,”:”). We also identify three inferred relations: IsSubClauseTo, IsSameAs and IsTypeOf.

If a noun phrase consists of several sub NPs combined with ‘,’ or/and a coordinator, the noun phrase is kept joined, and all sub clauses are also given an inferred relation IsSubClauseTo to the main NP (see Figure 3). Each sub clause is also given the sequence number of the main NP and a sub sequence number in the order they appear in the main NP e.g. NP10(1). The objective with inferred relation IsSubClauseTo is to visualize list of NPs.

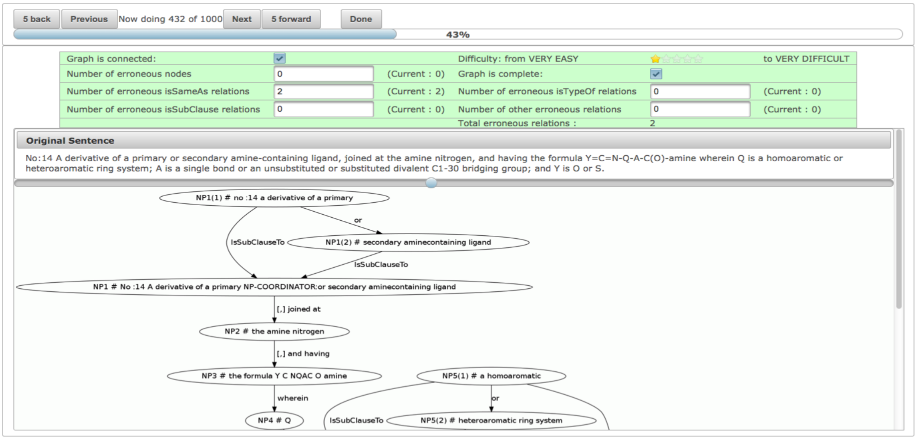
IsSameAs and IsTypeOf are partly associated with the discourse structure of the claims sentence. For IsSameAs, identifying the same entity, only the initial article/word may differ between ‘a’, ‘said’, or ‘the’ (see Figure 2). The relation IsTypeOf is assigned when one node has been pre-modified but the head noun of both NPs is the same as in Figure 2 where NP5 “low molecular weight chitin chitosan” IsTypeOf to NP3 “chitin chitosan”. The nodes and links are made into RDF triples, subsequently used in order to generate the sentence claims graphs[3].

3.2 Evaluation

To the best of our knowledge, there is no existing gold standard for any type of linguistic annotated information in the patent domain. Therefore, the assessment was based on manually assessing all graphs. Due to the fact that there are very few people having the level of deep linguistic knowledge, as well as the domain specific knowledge required to conduct assessment of linguistic accuracy of the displayed graphs, we decided upon a more generic evaluation schema.

The assessor group was divided into two groups: expert (3) and non-expert (14). The expert group consisted of linguists with some existing knowledge of the patent domain. The group of non-experts consisted of engineers and university students.

For the evaluation task, we constructed a simple interface showing the graph as well as the original sentence (see Figure 4).



Graph is connected:	☑		Difficulty: from VERY EASY	★☆☆☆☆	to VERY DIFFICULT
Number of erroneous nodes	0	(Current : 0)	Graph is complete:	☑	
Number of erroneous isSameAs relations	2	(Current : 2)	Number of erroneous isTypeOf relations	0	(Current : 0)
Number of erroneous isSubClause relations	0	(Current : 0)	Number of other erroneous relations	0	(Current : 0)
			Total erroneous relations :	2	

Original Sentence

No:14 A derivative of a primary or secondary amine-containing ligand, joined at the amine nitrogen, and having the formula Y=C=N-Q-A-C(O)-amine wherein Q is a homoaromatic or heteroaromatic ring system; A is a single bond or an unsubstituted or substituted divalent C1-30 bridging group; and Y is O or S.

NP1(1) # no :14 a derivative of a primary
or
IsSubClauseTo
NP1(2) # secondary aminecontaining ligand
IsSubClauseTo
NP1 # No :14 A derivative of a primary NP-COORDINATOR:or secondary aminecontaining ligand
[,] joined at
NP2 # the amine nitrogen
[,] and having
NP3 # the formula Y C NQAC O amine
NP5(1) # a homoaromatic
wherein
or
NP4 # Q
IsSubClauseTo
NP5(2) # heteroaromatic ring system

Fig. 4. Evaluation Tool

[3] Software: Graphviz (`http://search.cpan.org/~rsavage/GraphViz-2.14/lib/GraphViz.pm`)

Each claim graph was randomly given to one expert and one non-expert. Since the task turned out to be very time consuming, not all of the assessors assessed all of the claim graphs. All non-experts were assigned 100 claim graphs.

The assessment task consists of:

i) assessing the completeness of each displayed sentence claim graph where each word in the original sentence should also be represented in the graph.
ii) assessing connection, i.e. the graph must consist of a single connected component (ignoring edge directions).
iii) assessing if the claim graph displays correctly identified nodes and relations.

We defined seven parameters we asked the assessors to assess for each graph:

1. graph is complete,
2. graph is connected,
3. number of erroneous nodes,
4. number of erroneous IsSameAs relations,
5. number of erroneous IsSubClauseTo relation,
6. number of erroneous IsTypeOf relations,
7. number of other erroneous relations.

In the instructions for the evaluation task, a simple example and a domain example were given for each type of relation and erroneous nodes and relations. In order to find out how difficult assessors found the tasks to be, we asked each assessor to grade each graph from as scale 1 (very easy) to 5 (very difficult).

4 Results

We computed the inter-annotation agreement between expert and non-expert users for each assessment parameter. Similarly to [10], we also computed the inter-annotator agreement as the percentage of equal assessments among all pairs of assessments (Table 3).

Table 3. Inter-annotation agreement

Assessor Pair	No of sentences	Connected Graphs	Erroneous Nodes	Erroneous IsSameAs	Erroneous IsTypeOf	Erroneous IsSub-ClauseTo	Erroneous Other Relations	Complete graphs	Graph Difficulty
Non-expert vs Expert	182	98.35	68.13	87.91	97.80	96.15	69.78	84.62	26.37
Expert vs Expert	193	97.41	61.14	84.97	97.93	98.45	64.77	74.09	56.48

As seen in Table 3, the inter-annotation agreement of grading how difficult the task was for each graph was low between experts and non-experts (26.37%), while the experts agreed 56.48% on the difficulty associated with the task. The disagreement on difficulty does not come as a surprise, due to the wide range of options (1-5) as well as due to the graphs' complex structures, some of them containing up to 26 nodes and relations. Linguists are more trained to this type of visualization of sentences, which makes the interpretation and the reading of the graph more straightforward.

Assessing if the graph was connected, i.e. if for any pair of nodes in the graph there exists a path between them (potentially ignoring edge directions), was shown to be the easiest task when comparing inter-annotation agreement between pairs of non-expert vs expert users (98.35%) and expert vs expert (97.41%) users. For the other parameter assessments, the inter-annotation agreement between the pairs differs approximately by 10 percentage units. For the Erroneous Nodes and Erroneous Other Relations, there is a considerable drop in the agreement between both pairs as seen in Table 3. Compared to the other parameters assessed, these are by far the most loosely defined for the non-expert user. Furthermore, these two are strongly correlated – it is often the case that when one node is erroneously identified, an erroneous relation (other than IsSameAs, IsTypeOf, or IsSubClauseTo) is added to the graph.

As seen in Table 4 out of the 600[4] graphs evaluated most of the graphs are connected (for all assessors 97%, for only expert assessors 97%) and complete (for all assessors 86%, for only expert assessors 81%). Among the three inferred relations, IsSubClauseTo and IsTypeOf had the fewest errors. The algorithm extracting the IsSameAs relation was mistakenly identifying split chemicals as same units as well as identifying a longer sequence node with a smaller node containing a sub-sequence of the node.

Table 4. Assessment of evaluation parameters averaged over all assessors

IPC	# Sentences	Erroneous Nodes	Erroneous IsSameAs	Erroneous IsTypeOf	Erroneous IsSubClauseTo	Erroneous Other Relations	Complete Graph	Connected Graph	Difficulty
A	291	0.05 (0.10)	0.03 (0.07)	0 (0.02)	0.01 (0.04)	0.03 (0.09)	0.87 (0.31)	0.98 (0.13)	2.08 (1.16)
B	277	0.08 (0.11)	0.02 (0.06)	0 (0.01)	0 (0.01)	0.06 (0.10)	0.86 (0.32)	0.98 (0.14)	2.03 (1.14)
C	284	0.07 (0.10)	0.02 (0.07)	0 (0.01)	0 (0.03)	0.05 (0.09)	0.86 (0.32)	0.97 (0.16)	1.99 (1.13)
D	52	0.04 (0.08)	0 (0.02)	0 (0.01)	0 (0.02)	0.02 (0.06)	0.93 (0.24)	1 (0)	1.86 (1.04)
F	43	0.09 (0.09)	0.02 (0.05)	0 (0)	0 (0)	0.05 (0.08)	0.81 (0.33)	0.97 (0.17)	1.93 (0.82)
G	163	0.1 (0.12)	0.03 (0.07)	0 (0.01)	0.01 (0.03)	0.07 (0.10)	0.85 (0.34)	0.96 (0.18)	2.04 (1.16)
H	102	0.09 (0.10)	0.03 (0.06)	0.01 (0.02)	0 (0.01)	0.06 (0.09)	0.79 (0.39)	0.93 (0.25)	2.38 (1.31)
Total	1212	0.07 (0.11)	0.02 (0.07)	0 (0.02)	0 (0.03)	0.05 (0.09)	0.86 (0.32)	0.97 (0.16)	2.05 (1.15)

The Erroneous Nodes and Erroneous Other Relations is an indication when a noun phrase boundary has been wrongly identified. There were several where part of the relation between NP1 and NP2 "according to claim" contains the word claim that should have been part of the NP2 i.e. [A method] according to [claim 2]. The error

[4] The 1212 figure that appears in the table is due to the fact that many claims are assigned to more than 1 IPC Section.

was caused by the PoS tagger assigning the word “claim” the VB (verb) tag, and therefore affecting the chunker and ultimately the graph. This was triggered by “to”, which functions as an infinitive marker. The annotation of the word “claim” by the PoS-tagger in the sequence “according to claim” was unstable, randomly assigning VB (verb) or NN (noun). Only when the word “claim” was written with a Capital letter, the PoS-tagger identifies “Claim” as a Proper noun (NNP) which, given the context, is the most proper PoS-tag.

Furthermore, due to the PoS-tagger assigning words such as the “hydrogenated/VBN” (verb, past participle) instead of JJ (adjective), both erroneous nodes and relations were generated. This error is one of the most common erroneous and unstable annotations of the PoS-tagger, which clearly affects the NP boundary detection. The chunker was unstable in chunking digits into units. For instance the chunker joined the first digits [Claim/NNP 1/CD ,/, 2/CD] but not the last “,/, or/CC [3/CD]”. This did not depend on whether a coordinator (CC) was present or not.

Among old patents, there are also some stylistic text representations, which severely damage the text processing tools. For instance, it is common that the word characterized is written with space in between the letter sequence (“c h a r a c t e r i z e d”). This makes the entire graph more or less erroneous. Moreover, the presence of OCR-errors affected the extraction, e.g. the word ‘in’ mistakenly being identified as the letter ‘m’.

5 Conclusion

In this paper, we examined:

i) visualization of patent claim sentences in order to create a system which supports the cognitive process of analyzing patent claims
ii) domain adaptation of the NLP tools used in the pipeline, as well as a generic evaluation schema using non-experts and experts.

Our result shows that approximately 90% of all graphs used in the test collection have been assessed both by expert and non-expert to be complete, connected and having correctly identified nodes and relations. When comparing the inter-annotation agreement for each of the pre-defined erroneous nodes and relations we see that the expert-expert and the expert-non-expert values are similar. Consequently, the inter-annotation agreement indicates that for the pre-defined evaluation scheme, as presented in this study, using non-expert assessors is at least feasible and not as problematic as anticipated, despite the highly specific characteristics of the patent claims.

Our finding confirms that performance decreases when using existing general NLP tools when changing text focus from the mainstream genre text towards a specific text genre. Even if we used a state-of-the-art PoS-tagger for English with high accuracy, even small errors affected the parsing of a patent claims sentence negatively. Despite this, our result indicates that a general PoS-tagger and chunker can be used successfully on patent claims if combined with rules based upon observed syntactic patterns from the patent genre. In order to make the PoS-tagger even more robust when parsing patent text, a more extensive normalization procedure needs to be implemented dealing with chemical compounds, formulae and OCR-errors, as well as rejoining words written with spaces between their letters.

The general PoS-tagger used in this experiment still made errors, which caused the chunker to generate incorrect phrase boundaries and thereby caused the entire claim graphs to collapse. The average TTR values for each technical field (defined by the IPC section) suggest an alternation of the words distribution for claim sentences. The PoS-tagger can be made more robust by adding more post heuristic rules addressing complex noun phrase constructions. In the future, we will also investigate if post contra rules could be used as an intermediate layer in order to improve the performance of a full-scale parser in the patent domain. The method used to establish the DCG representation could be adjusted to intermediate layers in technical terminology extraction, as well as computing sub graph similarity.

To summarize, in this paper we have presented a DCG construction method applicable to all technical fields of the patent domain. We note that this is particularly important, since a support tool for patent experts needs to be able to deal with variations in terminology and linguistic features. We have also provided the experimental evidence to show that the tool achieves high success rates in identifying the important elements of an invention described in the claims, and the relations that bind them.

References

[1] Sheremetyeva, S.: Natural language analysis of patent claims. In: Proc ACL-2003, Workshop on Patent Corpus Processing, pp. 66–73 (2003)

[2] Hunt, D., Nguyen, L., Rodgers, M.: Patent Searching Tools & Techniques. John Wiley &Sons, New Jersey (2007)

[3] Lupu, M., Huang, J., Zhu, J.: Evaluation of Chemical Information Retrieval Tools. In: Croft, W.B., Lupu, M., Mayer, K., Tait, J., Trippe, J.A. (eds.) Current Challenges in patent Information Retrieval. Springer (2011)

[4] Hansen, P.: Task-based Information Seeking and Retrieval in the Patent Domain: Processes and Relationships. Tampere University Press (Doctoral dissertation), Tampere (2011)

[5] Uematsu, S., Kim, J.-D., Sujii, J.: Bridging the gap between domain-oriented and linguistically-oriented semantics. In: Proc ACL-2009, Workshop BioNLP 2009, pp. 162–170 (2009)

[6] Giesbrecht, E., Evert, S.: Part-of-speech tagging - A solved task? An evaluation of POS taggers for the Web as corpus. In: Alegria, I., Leturia, I., Sharoff, S. (eds.) WAC5 (2009)

[7] Ferraro, G.: Towards deep content extraction from specialized discourse: The case of verbal relation in patent claims Department of Information and communication Technologies: Universitat Pompeu Fabra (Doctoral dissertation) (2012)

[8] Parapatics, P., Dittenbach, M.: Patent Claim Decomposition for Improved Information Extraction. In: Lupu, M., Mayer, K., Tait, J., Trippe, J.A. (eds.) Current Challenges in patent Information Retrieval. Springer (2011)

[9] Ferraro, G., Wanner, L.: Towards the derivation of verbal content relations from patent claims using deep syntactic structures. Knowledge-Based Systems 24(8), 1233–1244 (2011)

[10] Verberne, S., D'hondt, E., Oostdijk, N., Koster, C.: Quantifying the Challenges in Parsing Patent Claims. In: Workshop of AsPIRe, pp. 14–21 (2010)

[11] Wäschle, K., Riezler, S.: Analyzing parallelism and domain similarities in the MAREC patent corpus. In: Salampasis, M., Larsen, B. (eds.) IRFC 2012. LNCS, vol. 7356, pp. 12–27. Springer, Heidelberg (2012)

[12] Koster, H.-A.C., Beney, J., Verberne, S., Vogel, M.: Phrase-Based Documentation Categorization. In: Croft, W.B., Lupu, M., Mayer, K., Tait, J., Trippe, J.A. (eds.) Current Challenges in patent Information Retrieval. Springer (2011)

[13] Justeson, S.J., Katz, M.S.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering 1(1) (1995)

[14] Bouayad-Agha, N., Casamayor, G., Ferraro, G., Wanner, L.: Simplification of Patent Claim Sentences for their Paraphrasing and Summarization. In: Lane, H.C., Guesgen, H.W. (eds.) The 22nd International Florida Artificial Intelligence Research Society Conference, Sanibel Island, Florida, USA, May 19-21, AAAI Press (2009)

[15] Shinmori, A., Okumura, M., Marukawa, Y., Iwayama, M.: Patent claim processing for readability: structure analysis and term explanation. In: Proc. ACL-2003 Workshop on Patent Corpus Processing, Stroudsburg, PA, USA, vol. 20, pp. 56–65 (2003)

[16] Andersson, L., Mahdabi, P., Hanbury, A., Rauber, A.: Exploring patent passage retrieval using nouns phrases. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp.
676–679. Springer, Heidelberg (2013)

[17] Ramshaw, A.L., Marcu, P.M.: Text Chunking Using Transformation-Based Learning. In: 3rd Workshop on Very Large Corpora, Cambridge, MA, USA (1995)

[18] Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proc. of HLT-NAACL, pp. 252–259 (2003)