Assessing four automatic term recognition methods: Are they domain-dependent?

Maria Jose Marin Perez, Camino Rea Rizzo

Abstract

Getting to know the terms in a specialised text certainly contributes to the understanding of the text itself. Their identification becomes essential precisely because of that reason and, owing to the large size of specialised corpora nowadays, the use of automatic term recognition (ATR) methods is fundamental when trying to extract the most characteristic terms in a given domain. However, these methods are not 100% effective and they must be validated before resorting to them so that the precision levels achieved are high enough for specialists to draw reliable conclusions on this type of vocabulary.

This article presents the assessment of four different ATR methods on two specialised corpora of legal and telecommunication English. The methods selected, TF-IDF (term frequency-inverse document frequency), C-value (Frantzi and Ananiadou, 1999) TermoStat (Drouin, 2003) and Terminus 2.0 (Nazar and Cabré, 2012) were evaluated in terms of precision. The aim of this evaluation is to compare the results obtained in all cases and to conclude whether there exists a certain degree of domain-dependence as regards each of these methods.

Keywords: automatic term recognition; legal English; telecommunication English; specialised corpora

Introduction

The lexical level in specialised texts is characterised, amongst other features, by the presence of terms which designate concepts and notions specific to a subject field of human activity, that is, the terminology of the discipline. Terms crystallise the knowledge shared by the specialised community. Spasic et al. (2005: 240) highlight this idea defining terms as "a textual realisation of a specialised concept". According to Cabré (2000: 62) terms are "*unidades de forma y contenido que, utilizados en determinadas condiciones discursivas, adquieren un valor especializado*". Hence, their recognition and extraction is fundamental for a better understanding of this type of texts. Furthermore, Nation (2001) and Nation and Waring (1997) underline the relevance of terms which cover 5% of the running words in any specialised text.

Specialised corpora are large collections of texts belonging to a given domain where, due to their size, term recognition becomes an unattainable task if it cannot be carried out in an automatic way. Thus, the use of effective automatic term recognition (ATR) methods is essential to fulfil this function. Literature reviews on ATR methods (Maynard and Ananiadou, 2000; Cabré et al., 2001; Lemay et al., 2005; Chung, 2003; Almela, 2008, etc.) show the great amount of techniques and procedures employed to identify and extract these units, as will be shown in section 3.

Nevertheless, the literature devoted to the assessment of these methods in different domains is reduced. Few initiatives such as the one described in Mondary et al. (2012), which studies the influence of corpus size and type on the efficiency of automatic term recognition, go along these lines. Other authors like Bernier-Colborne (2012: 1) show their concern about a lack of standard in ATR validation which is often carried out manually or employing a gold standard without being systematically described.

That is the reason why this article presents the validation of four different ATR methods applied on two specialised corpora of legal and telecommunication English with the aim of comparing the results obtained and drawing conclusions on their possible domain-dependency. The methods selected for evaluation are: TF-IDF (term frequency-inverse document frequency); TermoStat (Drouin, 2003); C-Value (Frantzi and Anniadou, 1999) and Terminus (Nazar and Cabré, 2012).

Section 2 concentrates on the description and justification of both the corpora employed in this study, Telematics Corpus (TC) and United Kingdom Supreme Court Corpus (UKSCC). Section 3 focuses on the review of the literature on ATR methods and the description of the four methods selected for evaluation. Finally, section 4 deals with the implementation of these methods and the analysis of the results of the experiment. The conclusions drawn after the analysis of the data obtained are described in section 5.

Description of *TC* and *UKSCC:* two specialised corpora

The amount and availability of specialised corpora is reduced[1]. That is why, in order to have a reliable source of specialised vocabulary which could be employed to different purposes, two specialised English corpora were designed and compiled *ad hoc*. *TC* and *UKSCC* were created following the standards for general corpus design and compilation in Sánchez et al. (1995) and those for specific corpora in Pearson (1998) and Rea (2010).

The Telematics Corpus is actually a subcorpus belonging to a main corpus specialized in Telecommunication Engineering English (TEC) (Rea, 2010). Telematics is one branch of telecommunications which has been selected for comparison purposes. TEC is a fairly representative sample of 5.5 million words of academic and professional

written English extracted from a wide range of sources (magazines, books, web pages, journals, brochures, advertisements and technology news), originating in native and non-native parts of the world and covering 18 subject areas subsumed under seven major areas of knowledge (Electronics; Computing Architecture and Technology; Telematic Engineering; Communication and Signal Theory; Materials Science; Business Management; and System Engineering) and two branches of expertise in telecommunication engineering (Communication Networks and Systems; and Communication Planning and Management). All the language samples were produced in communicative acts where at least one of the speakers was a professional or expert in the domain.

The selected subcorpus defines the main area of Telematics which deals with telecommunication network operations and covers four subject domains (Telematics, Communication Networks and Services, Telecommunication Systems and Switching). The whole of samples adds up to 1.2 million running words.

UKSCC, in turn, is a legal corpus of law reports (written collections of judicial decisions) of 2.6 million-words. The reasons to focus on this genre to study the linguistic properties of legal terminology are varied. To begin with, the UK belongs to the realm of common law, as opposed to civil or continental law, which is the judicial system working in most Western European countries. In purely common law systems, the acts passed at their parliaments have gained greater importance being most often cited in case decisions. However, case law stands at the very basis of common law systems which rely on the principle of binding precedent to work, that is to say, a case judged at a higher court must be cited and applied whenever it is similar to the one being heard in its essence (the *ratio dicendi*), and judicial decisions are employed by law practitioners as the basis for their arguments, decisions, etc.

Another fact that makes law reports an outstanding legal genre is that they not only cover all the branches of law, but might also present full embedded sections of other public and private law genres displaying therefore great lexical richness and variety. Following Sinclair (2005: 5) "the contents of the corpus should be selected … according to their communicative function in the community in which they arise". Consequently, such texts as these have been chosen to form the corpus due to the pivotal role they play in common law legal systems. The Supreme Court was selected as the text source owing to its relevance within the British judicial system (all the decisions made at the Supreme Court set precedent and are cited whenever applicable), and the wide lexical variety of the documents coming from it. It is at the top of the UK judicial pyramid and deals with cases belonging to all branches of law. As for its structure, UKSCC is a synchronic, monolingual and specialised collection of 193 judicial decisions from the UK Supreme Court and the House of Lords[2] issued between 2008 and 2010. The documents included in UKSCC are authentic judicial decisions as produced by this legal institution in raw text format.

ATR methods review and description

Literature review

The literature reviews devoted to the study of ATR methods are numerous and usually group them depending on the parameters considered for the identification and extraction of candidate terms. On the one, hand there are purely statistical methods such as Church and Hanks (1990), Ahmad et al. (1994), Nagakawa and Mori (2002), Chung (2003a), Fahmi et al. (2007), Scott (2008) or Kit and Liu (2008), to name but a few. On the other hand, there exist methods which concentrate solely on linguistic data, namely,

Ananiadou (1988), David and Plante (1990), Bourigault (1992) or Dagan and Church (1994). Finally, other ATR methods rely on the combination of both statistical and linguistic data. The work of Justeson and Katz (1995), Daille (1996), Frantzi and Ananiadou (1996; 1999), Jaquemin (2001), Drouin (2003), Barrón Cedeño et al. (2009) or Nazar and Cabré (2012) illustrate this trend, amongst others.

Method description

The methods described in this section were selected due to their varied nature. While TermoStat (Drouin, 2003) and TF-IDF only identify single-word terms (SWTs)[3], C-value (Frantzi and Ananiadou, 1999) manages to recognise multi-word terms (MWTs) and Terminus (Nazar and Cabré, 2012) can extract both. On the other hand, TermoStat and Terminus resort to corpus comparison needing a reference corpus of general English to work, whereas TF-IDF and C-value do not require such procedure. Moreover, TermoStat and Terminus are online term extraction tools which carry out the whole process in an automatic way while the algorithms corresponding to TF-IDF and C-value had to be implemented either manually or using other tools to make them work, as will be described in detail below. Let us then concentrate on the description of these four methods before analysing the results obtained after implementing them on both corpora.

To begin with, TF-IDF (term frequency-inverse document frequency) relies on two main parameters obtained from the analysis of one single corpus. It is a purely statistical method since it takes into account a word's frequency in the corpus and also the number of documents/texts the word occurs in throughout the whole document collection. The

higher a word's frequency in the corpus and the fewer documents it appears in, the greater weight it will display.

IDF was originally proposed by Sparck Jones (1972). She believed that, contrary to the general belief that a word's level of representativeness was directly related to its occurring in many texts within a corpus, a word appearing in fewer documents might potentially be more representative within a given document collection.

TF-IDF is the result of multiplying IDF, which is "defined as $-\log_2 df_w/D$, where *D* is is the number of documents in the collection and df_w is the document frequency, the number of documents that contain [the word] *w"* (Church and Gale, 1995: 121), by a word's frequency in a given document (TF).

In the present article, the IDF formula implemented on both the legal and telematic corpora is the one proposed originally by Sparck Jones. Nevertheless, the parameter TF was adapted for the sake of comparison with the lists produced by the other four methods. Instead of resorting to the frequency of a word within a single document, after calculating a word's IDF value, it was multiplied by the normalised frequency value of that word in the whole corpus.

Termostat (Drouin, 2003) is the second SWT recognition method validated in this study. Unlike TF-IDF, Drouin's method is fully automatic and can be easily implemented online using a free term extraction tool[4]. This method offers the possibility of identifying both SWTs and MWTs although, in this case, it was configured to focus solely on the former. Termostat offers the possibility of processing texts in French, English, Spanish, Italian and Portuguese of up to 30Mb in raw text format.

It is a hybrid method based on previous work on lexicon specificity such as Muller's (as cit. in Drouin, 2003: 100), Lafon's (ibid.), or Lebart and Salem's (ibid.). Drouin claims that the frequency of technical terms in a specialised context differs, in one way or

other, from the same value in a general environment and that "focusing on the context surrounding the lexical items that adopt a highly specific behaviour (...) can help us identify terms" (ibid.). The author uses a corpus comparison approach which provides information on a candidate term's standard normal distribution.

This ATR method uses Schmid's (1994, 1995) Tree Tagger as lemmatiser and POS tagger which leads to a list of candidate terms arranged according to their level of specificity. The system also offers other possibilities such as ranking them based on other measures like T-score, chi-square or log-likelihood which have not been assessed in the present article.

A threshold value of + 3.09, which acts as a cut-off point to discriminate terms from non-terms, is established to minimise the amount of noise (false positives). As it employs POS tagging, TermoStat can detect all lexical categories, namely, nouns, verbs, adjectives and adverbs although it can be configured to only focus on one of them depending on the user's preferences.

The author validates his method both automatically and resorting to three specialists. According to the judges' evaluation and the comparison with a gold standard (a telecommunication terminology database), TermoStat manages to identify 86% true terms on average. Nevertheless, Drouin insists, on the one hand, on the subjectivity of human validation processes where consensus is sometimes hard to reach, and, on the other hand, on the importance of complementing this type of methods with others that can help to study those words which activate a specialised meaning in a specific context.

On the other hand, *C-value* (Frantzi and Ananiadou, 1999) is a hybrid method which employs both linguistic and statistical data to produce a list of candidate terms ranked

according to their termhood score. A term's c-value can be calculated with respect to its frequency and the frequency of its sub-terms:

$$CValue(a) = \log_2|a| \cdot \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right)$$

Where, f (a) is the frequency of term (a) with $|a|$words, T_a is the set of candidate terms recognised by the method that contain (a) and $P(T_a)$ is the total number of longer candidate terms that contain (a).

The linguistic part of this method is articulated into different steps which go as follows:

1- The corpus is POS tagged.

2- A linguistic filter is applied so as to discard certain patterns and keep a balance between precision and recall (the use of an open filter could favour recall at the expense of precision). Only those strings containing nouns premodified by other nouns, adjectives or combinations of both are kept.

3- A stop list is employed which comprises both function words and high frequency ones from a sample corpus not expected to be terms.

As part of the statistical parameters utilised to select the candidate terms, the authors take into consideration the frequency of occurrence of the pattern, also the frequency of the pattern as part of other longer structures, the amount of these longer structures and the number of constituents of the pattern.

Frantzi and Ananiadou introduce the concept of 'nested terms' as key within the statistical part of their method. With the purpose of trying to discard those patterns which are not true terms, they decide to select only those which contain strings which also appear by themselves in the corpus displaying relatively high frequency. A frequency threshold of >3 is applied to avoid producing a too long list that might become a hindrance for the experts evaluating the output.

For the assessment of their method, the authors highlight the fact that there is no agreement amongst experts and that such subjectivity necessarily leads to the introduction of the concept of 'relative' precision and recall. Instead of asking an expert to extract all the terms in a corpus, which is time-consuming and hard to attain, recall figures are obtained "with respect to frequency of occurrence, which we use as the baseline method" (1999: 8).

The authors also assess precision at three stages: first, evaluating those candidates which have appeared as nested; second, evaluating only those appearing as nested and third, evaluating all the candidate terms. As a result, the authors realise that, in general, the use of a more open linguistic filter does not affect precision significantly. Moreover, using other statistical data "apart from the pure frequency of occurrence of candidate terms, improves the precision of the extracted nested multi-word terms, with a slight only loss on recall" (ibid. 13).

Finally, Nazar and Cabré (2012) propose an ATR method, freely available online[5], where term extraction becomes a fast and easy task. *Terminus 2.0* offers different possibilities for the researcher working with specialised terminology. As indicated on the website guide, it has varied functions such as textual corpus search, compilation and analysis, term extraction, glossary and project management, database creation and maintenance, and dictionary edition.

Their ATR method is based on the assumption that the system can learn how to recognise terms based on the language samples provided by the user. The expert does not need to formulate rules to help the system work but rather let it learn from the real samples provided of both specialised terms and general language using the latter for comparison.

The program "develops a statistical model with an abstraction of the main characteristics of both samples" (ibid. 210). As it is open to any user who can upload glossaries and corpora to help the system learn to identify terms in different domains, the more users employ it, the greater its capability will become to identify terminological units. As stated by the authors, the greatest innovation of this method is its collaborative character since it "allows a community of terminologists to share knowledge acquired by the program in each training phase ... As a consequence, our program is constantly improving in both precision and recall, as a sort of lifelong learning algorithm" (ibid. 212).

The method applied by the system is structured into three distinct phases: syntactic, lexical and morphological. To begin with, using Schmid's (1994, 1995) *Tree Tagger*, the texts are POS tagged and a syntactic model is developed based on the frequency of distribution of the syntactic patterns identified. After doing so, the frequency of the lexical units within those patterns is measured. Finally, it extracts initial and final character *n*-grams. The termhood score is obtained by assigning a higher value to those units which have a "significant frequency in the LSP training material with respect to the general language corpus" (Nazar and Cabré, 2012: 212). This process is followed for all levels of training.

The authors act as judges to validate their method by confirming the candidates extracted as true terms and discarding those which do not qualify as such. The corpus employed as the training set is a 300,000 word collection of papers on corpus linguistics published in 2010. The test corpus is also a collection of papers on the same topic of similar size (340,000 words). Both sets of texts were taken from the scientific journal Computational Linguistics. The reference corpus consists in a 2 million-word collection of press articles from the Leipzig Corpora Collection (as cit. in ibid.). In the evaluation

process the algorithm is trained also using *n*-gram frequency lists and word association measures.

As part of this training, the authors validate 800 terminological units and train the algorithm using this list of terms (both SWTs and MWTs). Once the training phase is accomplished, the study corpus is processed employing the information derived from the training. For the validation of the results obtained after processing the study corpus of 340,000 words, the authors resort to three different classical measures, namely, chi-square test, mutual information and frequency (the most frequent 1500 bigrams are extracted). They also employ a stop word list to filter the results.

As a result, the precision levels achieved are considerably better than those attained by the three methods used for comparison. Terminus is reported to attain 85% precision for the top 200 candidates and 75% for the top 400.

Method implementation and results

Method implementation and validation procedure

As regards their implementation, both Drouin's (2003) TermoStat and Nazar and Cabré's (2012) Terminus 2.0 were applied automatically to both UKSCC and TC. Both of them only require registering on their websites. Once registered and logged on to the system, uploading and processing both corpora just took a few minutes. The results were arranged according to the candidate's specifity score (for TermoStat) and its weight (for Terminus). As they require lemmatisation to calculate a word's termhood level, both methods provide a list of the lemmas and also their variants as well as many

other options which have not been considered for this study since they do not affect precision.

Conversely, calculating the values corresponding to TF-IDF and C-value (1999) was more complex due to the fact that the actual algorithms proposed by the authors had to be implemented semi-automatically on the list of word types obtained prior to their processing. This word type list was produced using Scott's (2008a) Wordsmith 5.0 software so as to have all the necessary information to carry out such process. The list was filtered to eliminate function words from it.

Wordsmith provides the data related to the raw frequency of the types in a corpus and also to their document frequency, that is, not only does it inform us about how many times a type repeats itself throughout the document collection, but also how many documents it occurs in, amongst other data. Therefore, applying Sparck Jones' formula to obtain IDF and then multiplying it by the word's frequency was a relatively simple task.

Nevertheless, C-value required more information and its implementation was easier thanks to JATE tools, the online[6] java tool set designed by Zhang et al. (2008) which allows the user to process corpora and implement eight different state-of-the-art ATR algorithms.

These four methods were assessed in terms of the precision levels achieved by each of them, that is, how many true terms were identified by each method with respect to the whole list of candidates. As the size of the output lists varied considerably (Drouin's method establishes a threshold which led to a shorter inventory of candidates, whereas the TF-IDF list included all the word types generated by Wordsmith), the total number of candidates considered for evaluation was 1,400 so that a similar assessment process could be followed in all cases.

The validation of the four methods selected was carried out both manually and employing two specialised glossaries as gold standard. The authors, feeling confident enough in both legal and telecommunication English, acted as judges checking the lists once they had been validated automatically to reduce silence (undetected terms) to the minimum. This manual supervision led to 3/4% improvement as regards precision.

The automatic validation of the lists was performed by resorting to two specialised electronic glossaries of legal English (of 10,054 terms) and telecommunication English (of 5,102). The comparison was done using an excel spreadsheet that would facilitate the identification of the true terms present in each of the lists under evaluation.

Results

The results of such comparison and later supervision led to the calculation of both average and cumulative precision, as illustrated in the graphs below.

Figure 1 shows the levels of precision attained on *UKSCC*, the legal corpus, by the four ATR methods for each group of 200 candidate terms from the list of 1,400 evaluated, which were arranged according to their termhood level[1] for each method (tables 1 and 2 below illustrate the top 25 candidate terms extracted by each method from both corpora).

[1] Authors vary in the way they refer to a word's level of specialisation so *termhood* is employed here to refer to that value.

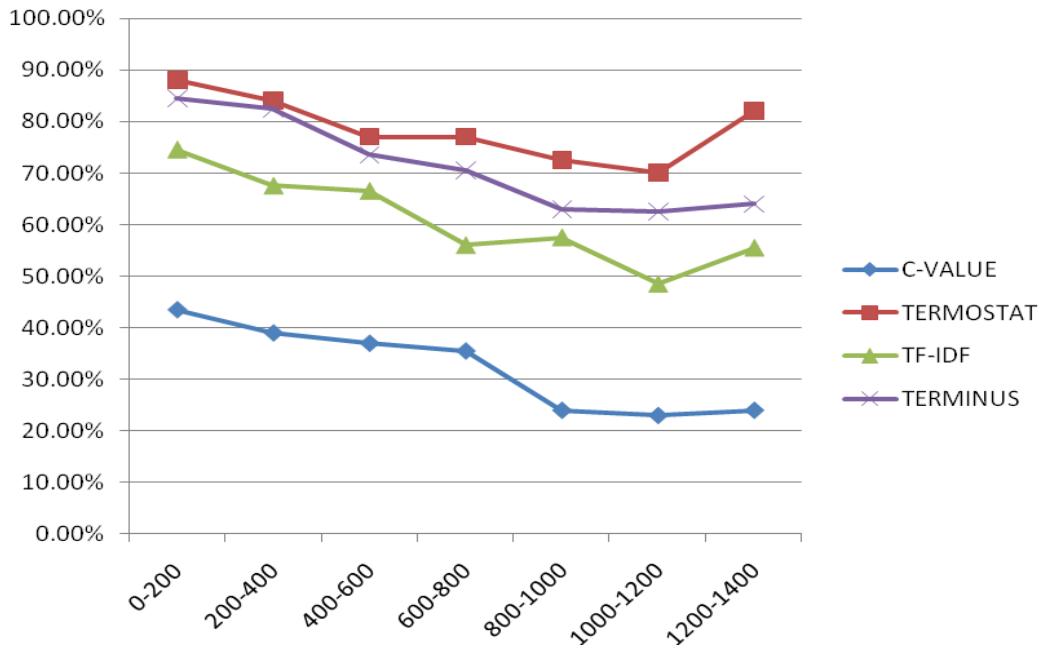

Fig 1 *Cumulative precision achieved on top 1400 candidates extracted from UKSCC*

Drouin's (2003) *TermoStat* is the most efficient method as it manages to identify 79% SWTs on average reaching a peak of 88% for the top 200 candidates in the list. It descends constantly to 70% precision from candidates 1000 to 1200 although it climbs up again to 82% by the end of the graph. *Terminus* stands in second position behaving in a very similar way. It meets its precision peak at 84.50% for the top 200 candidate single and multi-word terms and progressively descends to 64% for candidates 1200 to 1400. On average, TF-IDF is less effective. Standing in third position, it recognises 60.86% true terms as a mean value remaining 18 and 11 below the other two methods respectively.

Finally, *C-value* is the worst performing method whose efficiency in identifying MWTs is rather low. It is far below the rest of methods at 38 points on average, only reaching 43.50% precision on the top 200 MWT candidates.

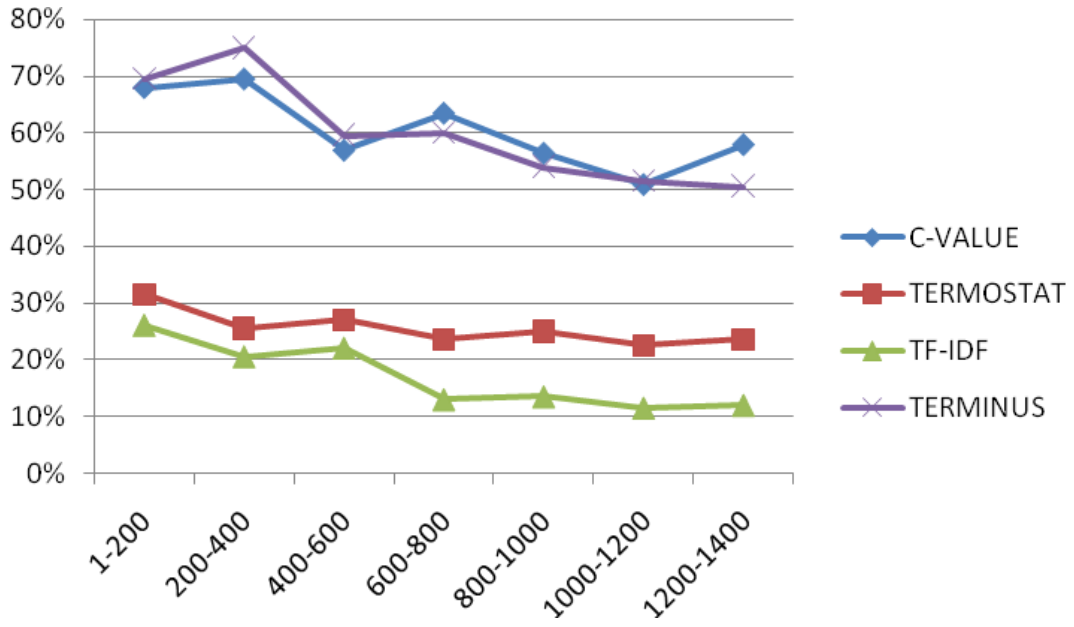

Fig. 2 *Cumulative precision achieved on top 1400 candidates extracted from TC*

As shown in figure 2, the ATR methods evaluated herein, in general, produce worse results when applied to the telematics corpus than to the legal one. *Terminus* (Nazar and Cabré, 2012) and *C-value* (Frantzi et al. 1999) are the most efficient methods reaching 60% and 61% precision respectively on average. From candidates 1 to 600, *Terminus* manages to identify 68% true terms while *C-value* extracts 65% within the same range. From that point on, *C-value* outperforms *Terminus* standing 3.25 point above it until the end of the graph.

As opposed to the results obtained with the legal corpus, Drouin's (2003) *TermoStat* does not manage to extract more than 24.36% true terms in *TC*, the telematics corpus. It stands in third position at almost 7 points above TF-IDF, the fourth one. Both methods behave similarly and do not decrease their efficiency sharply keeping relatively constant until the end of the graph, especially *Termostat*, basically owing to their poor performance from the beginning of the list.

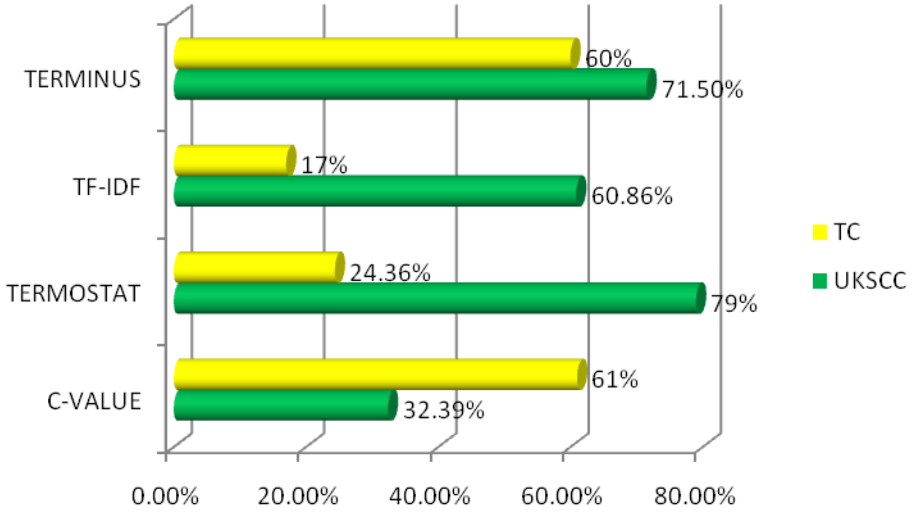

Fig. 3 Average precision attained by all methods on both UKSCC and TC.

Judging by average precision illustrated by figure 3 above, except for C-value, which is almost twice as precise when implemented on TC as it is on UKSCC, the rest of the methods assessed prove to have performed better when applied to the legal corpus than to the telematic one. The figures are particularly striking when it comes to TermoStat and TF-IDF which display a substantial difference of 55 and 43 points respectively between UKSCC and TC. Terminus is the only method whose efficiency does not differ so markedly between both corpora as there is a difference of 11 points. Thus, having analysed the results obtained after processing both corpora and implemented the four methods on each of them, could it be stated that the ATR methods validated in this study are domain-dependent?

As indicated by the graphs, Terminus is the only method which does not appear to be so linked to a specific domain as the other three since precision does not differ so sharply between both corpora. However, as regards TF-IDF, TermoStat, and C-value, the figures are striking showing that the first two perform much better within the legal field while precision is twice as high in the Telematics domain in the case of C-value.

Nevertheless, as already pointed out by Drouin (2003), human validation does impose a degree of subjectivity on validation processes and consensus is hard to achieve particularly concerning the vocabulary which is not highly specialised[7]. Furthermore, the conclusions and figures shown in this study might differ from those drawn by other specialists in a similar context.

In addition, Bernier-Colborne's (2012) concern about a lack of standard in the design and use of a gold standard (terminological databases or glossaries) which might affect the results of automatic validation must also be underlined. That is why the lists were manually checked by the authors to try and minimise both noise and silence caused not by the lack of efficiency of the ATR methods implemented, but by a potentially wrong design of the specialised glossaries employed for validation, inevitably adding a degree of subjectivity to that supervision. To conclude, tables 1 and 2 show the top 25 terms extracted by each method from both the legal and telematic corpora.

Table 1 List of top 25 candidate terms extracted from UKSCC.

DROUIN		**TF-IDF**		**C-VALUE**		**TERMINUS**	
Section	126.29	Land	0.998	United kingdom	1869.5859	Reasonable	397271.6251
V (versus)	112.55	Article	0.965	Noble and learned friend	1488.0410	Basis	299498.7355
Case	111.79	Contract	0.926	Human right	1059.1502	Extent	271501.0668
Para (paragraph)	108.63	Jewish	0.898	Lord hope	807.18059	Payment	243836.7974
Article	97.39	Extradition	0.866	Present case	789.88435	Lawful	235189.5609
Court	88.65	Possession	0.861	Common law	770.94536	Witness	230170.3331
Appeal	80.3	Child	0.845	Lord hoffmann	770.18334	Word	198149.7375
Appellant	78.47	Tenant	0.804	Learned friend	711.21424	Facie	191377.0993
Law	73.55	Company	0.783	Lord bingham	673.08374	Context	146508.8683
Judgment	71.67	Convention	0.775	Member state	587.42966	Payable	145321.5141
Claim	69.8	Asylum	0.724	Lord brown	569.71471	Causation	135029.1021

Right	67.98	Data	0.721	Local authority	553.85301	Injunction	121506.2169
Apply	65.5	Directive	0.702	Lord rodger	512.56357	Complaint	112844.924
Order	64.39	Equipment	0.701	Lord walker	494.12551	Obligation	112659.4599
Decision	63.53	Immigration	0.656	Public interest	437.61408	Infringement	101451.544
Person	62.83	Discrimination	0.647	No doubt	435.64845	Wording	93573.26866
Proceeding	61.7	Suicide	0.645	Judicial review	409.79260	Presumption	89657.53631
Relevant	59.02	Rent	0.645	Convention right	400.74187	Actual	89491.33949
Purpose	58.45	Accommodation	0.627	Northern Ireland	400.45303	Inference	88221.7819
Defendant	57.72	Planning	0.614	European court	397.11444	Lawfulness	85915.14383
Provision	57.55	Criminal	0.614	Strasbourg court	396.04405	Misconduct	84649.52406
Principle	55.77	Commissioners	0.608	Lord mance	391.76249	Judgment	55907.31426
Application	55.5	Clause	0.583	Home department	383.34534	Doctrine	54505.99677
Jurisdiction	55.5	Property	0.580	Lord phillips	360.55652	Easement	52735.42997
Paragraph	54.69	Lease	0.576	Public authority	357.99171	Suicide	51198.75113

Table 2 List of top 25candidate terms extracted from TC.

DROUIN		**TF-IDF**		**C-VALUE**		**TERMINUS**	
Network	147.15	LSAS	3204.4700	Project management	424.6557	Output	45153.6911
Router	114.79	LSA	3023.7244	Designated router	404.4089	ATM	41730.7813
User	109.23	Groupware	2477.5944	IP address	371.4250	Grouplet	27408.7903
Use	91.12	Linux	2124.9761	Frame relay	312.5854	Adjacency	23678.7352
Packet	87.13	Packet	2072.8506	Service provider	309.6958	Adjacency	22640.0821
Service	85.26	Scheme	2044.0293	Operating system	274.4109	Subnets	19364.0942
Interface	84.66	Packets	1988.3902	Hello packet	273.9527	Linux	18717.9836
Datum	83.92	Directory	1985.8265	Routing protocol	259.3149	TCP	17708.7487
Application	82.61	Program	1789.7763	Routing table	259.2958	Context	15734.6601
Protocol	80.97	Cell	1769.1162	IP VPN	219.1622	Topology	15356.2259
Server	79.63	Figure	1736.0456	Cell phone	209.8727	Encoder	14744.9678
System	77.18	ATM	1725.6936	Access manager	198.7741	Octet	14266.4384

Object	75.88	Objects	1700.8508	Web service	198.6145	Protocol packet	13137.0448
Program	73.02	Frame	1684.3549	Database description packet	196.2349	Database	12334.3132
Address	72.55	G	1681.7828	Groupware system	192.0679	Octet	10189.3372
Internet	72.3	Layer	1652.8426	Data structure	187.1699	Graph	10162.2551
Software	66.91	Ethernet	1641.1827	Backup designated router	178.8499	Protocol	9153.84327
Information	64.52	File	1640.5399	Secure server	152.8515	Byte	9034.5777
Link	63.73	Collaborative	1610.0374	Collaborative object group	150.0054	Bytecode	8663.8622
Routing	61.86	Procedure	1598.8956	Service component	147.5643	Wavelength	8635.3268
Model	60.21	VPN	1575.9871	User interface	145.7148	ATM network	8222.8878
Type	59.62	B	1565.7557	Metro ethernet access service	145.3435	Text	8105.9186
Traffic	59.43	MPLS	1563.6814	Eliteconnect wlan security system	141.3627	Expression	8013.8077
File	59.3	Database	1543.0457	BGP MPLS VPN	141.1912	Iteration	7864.0932
Define	59.18	Neighbor	1534.0951	Remote object	138.8357	Browser	7668.4494

Conclusion

This article has presented the assessment of four different ATR methods on two specialised corpora of legal and telecommunication English as regards the precision levels achieved by each of them. After describing the corpora employed in this experiment, the methods singled-out for evaluation and their process of implementation, the data obtained have been discussed.

Except for one of the methods, C-value, which is twice as efficient in the telematic domain as it is in the legal one (61% against 32.39% respectively), the other three methods prove to perform worse in the telematic field. The differences are particularly noticeable for TermoStat which only reaches 24.36% precision in the telematic domain

while it manages to identify 79% true terms in the legal field. The precision levels attained by TF-IDF are lower in both cases displaying a difference of 43 points between both corpora. Terminus is the only method which, in spite of being less effective in the telecommunication realm, does not appear to be so closely linked to the domain it is applied to since the difference shown between both corpora is 11 points. Therefore, it can be stated that, as far as the corpora employed in this experiment are concerned, C-value, TermoStat and TF-IDF are domain-dependent, whereas Terminus, although performing better in the legal field, does not appear to display such dependence so clearly.

Finally, it must be emphasised that there exists no standardised method to design and compile the glossaries or terminological databases used as gold standard and this fact, as highlighted by Bernier-Colborne (2012), might affect experiments like the one described herein. Moreover, the fact that the output lists produced by each method were supervised by the authors manually to reduce noise and silence may have also added a certain degree of subjectivity that could alter the results to a certain extent too. That is the reason why exploring the degree of subjectivity implied in human validation processes might be interesting to tackle as further research related to ATR method validation.

References

Ahmad, K., Davies, A., Fulford, H., Rogers, M. (1994). "What is a term? The semi-automatic extraction of terms from text" in Snell-Hornby, M., Pöchhacker, F. and Kaindl, K. (eds.), *Translation Studies: An Interdiscipline*, 267-278. Amsterdam: John Benjamins.

Almela, A. (2008). *Evaluating Multiword Automatic Term Recognition Techniques on a Veterinary Medicine Corpus.* MA Thesis. Murcia: Universidad de Murcia.

Ananiadou, S. (1988). *A Methodology for Automatic Term Recognition.* PhD Thesis, University of Manchester Institute of Science and Technology: United Kingdom.

Barrón-Cedeño, A., Sierra, G. E., Drouin, P. and Ananiadou, S. (2009). "An Improved Automatic Term Recognition Method for Spanish" in Gelbukh, A. (ed.) *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009),* 125-136. Springer.

Bernier-Colborne, G. (2012). "Defining a Gold standard for the evaluation of Term Extractors" in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* URL: http://www.lrec-conf.org/proceedings/lrec2012/index.html [20/12/2012].

Bourigault, D. (1992). "Surface grammatical analysis for the extraction of terminological noun phrases" in *Proceedings of the 5th International Conference on Computational Linguistics, COLING-92,* 977-81. Nantes. URL: http://acl.ldc.upenn.edu/C/C92/C92-3150.pdf [05/02/2013]

Cabré, M.T. (2000). "Terminologie et linguistique: la théorie des portes". *Terminologies nouvelles. Terminologie et diversité culturelle* 21: 10-15.

Cabré, M. T., Estopà, R., Vivaldi, J. (2001). "Automatic term detection: a review of current systems" in Bourigault, D., Jacquemin, C., L'Homme, M.C. (eds.), *Recent Advances in Computational Terminology,* 53-87. Amsterdam: John Benjamins.

Chung, T. M. (2003a). "A corpus comparison approach for terminology extraction". *Terminology* 9, 2: 221-246.

Chung, T. M. (2003b). *Identifying Technical Vocabulary*. PhD thesis. Victoria University of Wellington.

Church, K.W., and Hanks, P. (1990). "Word association norms, mutual information, and lexicography". *Computational Linguistics* 16, 1: 22-29.

Church, K.W., and Gale, W. (1995). "Inverse Document Frequency (IDF): A measure of Deviations from Poisson" in *Proceedings of the Third Workshop on Very Large Corpora*, 121-130. Cambridge: Massachusetts Institute of Technology.

Dagan , I. and Church, K. (1994). "TERMIGHT: Identifying and Translating Technical Terminology" in *4th Conference on Applied Natural Language Processing*. URL: http://www.aclweb.org/anthology-new/A/A94/A94-1006.pdf [05/02/2013].

Daille, B. (1996). "Study and implementation of combined techniques for automatic extraction of terminology" in Klavans, J.L., and Resnik, P. (eds.) *The Balancing act: Combining symbolic and statistical approaches to language.* Cambridge, MA: MIT Press.

David, S. and Plante, P. (1990). *Termino* 1.0. Research Report of Centre d'Analyse de Textes par Ordinateur. Université du Québec, Montréal.

Drouin, P. (2003). "Term extraction using non-technical corpora as a point of leverage". *Terminology* 9, 1: 99-117.

Fahmi, I., Bouma, G. and Van der Plas, L. (2007). "Improving statistical method using known terms for automatic term extraction" in *Computational Linguistics in the Netherlands-CLIN* 17: 1-8.

Frantzi, K.T. and Ananiadou, S. (1996). "Extracting nested collocations", in *Proceedings of the 16th Conference on Computational Linguistics* 1: 41-46.

Frantzi, K.T. and Ananiadou, S. (1999). "The c/nc value domain independent method formulti-word term extraction". *Journal of Natural Language Processing* 3, 2: 115-127.

Jacquemin, C. (2001). *Spotting and discovering terms through NLP*. Massachusetts: MIT Press.

Justeson, J.S. and Katz, S.M. (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text". *Natural Language Engineering* 1: 9-27.

Kit, C and Liu, X. (2008). "Measuring mono-word termhood by rank difference via corpus comparison". *Terminology* 14, 2: 204-229.

Lemay, C., LHomme, M.C., Drouin, P. (2005). "Two Methods for Extracting 'Specific' Single-word Terms from Specialised Corpora: Experimentation and Evaluation". *International Journal of Corpus Linguistics* 10, 2: 227-255.

Maynard, D. and Ananiadou, S. (2000). "TRUCKS: A model for automatic multi-word term recognition". *Journal of Natural Language Processing* 8, 1: 101–125.

Marín, M. J., Rea, C. (2011). "Design and compilation of a legal English corpus based on UK law reports: the process of making decisions", in Carrió Pastor, M.L. and Candel Mora, M.A. (eds.). *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora*. *Actas del III Congreso Internacional de Lingüística de Corpus* 101-110. Valencia: Universitat Politècnica de València.

Mondary, T., Nazarenko, A., Zargayouna, H., and Berreaux, S. (2012). "The Quaero Evaluation Initiative on Term Extraction" in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).*
URL: http://www.lrec-conf.org/proceedings/lrec2012/index.html [23/01/2013].

Nakagawa, H. and Mori, T. (2002). "A simple but powerful automatic term extraction method" in *COLING-02 on COMPUTERM*. *Proceedings of the Second International Workshop on Computational Terminology* 1-7.

Nation, P. and Waring R. (1997). "Vocabulary Size, Text Coverage and Word Lists" in Schmitt, N. and M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, 6-19. Cambridge: CUP.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nazar, R., Cabré, M.T. (2012). "Supervised Learning Algorithms Applied to Terminology Extraction" in Aguado de Cea, G., Suárez-Figueroa, M.C., García-Castro, R., Montiel-Ponsoda, E. (eds.), *Proceedings*

of the 10th Terminology and Knowledge Engineering Conference (TKE 2012). Madrid: Ontology Engineering Group, Association for Terminology and Knowledge Transfer.

Pearson, J. (1998). *Terms in Context*. Amsterdam: John Benjamins Publishing Company.

Rea, Camino. (2010). **"**Getting on with Corpus Compilation: from Theory to Practice". *ESP World*, 1 (27), vol. 9.

URL: http://www.esp-world.info/articles_27/camino%20rea.pdf [05/02/2013].

Sánchez, A. et al. (1995). *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL.

Schmid, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees" in *Proceedings of International Conference on New Methods in Language Processing*, 44-49. Manchester, UK.

Schmid, H. (1995). "Improvements in Part-of-Speech Tagging with an Application to German" in *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Scott, M. **(**2008a). *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Sinclair, J. (2005). "Corpus and Text: Basic Principles", in Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice.* AHDS Literature, Languages and Linguistics: University of Oxford.

URL: http://ota.ahds.ac.uk/documents/creating/dlc/index.htm [05/02/2013]

Sparck Jones, K. (1972). "A statistical interpretation of term specificity and its application in retrieval". *Journal of Documentation* 28: 11-21.

Spasic, I., Ananiadou, S., McNaught, J. and Kumar, A. (2005). "Text mining and ontologies in biomedicine: Making sense of raw text". *Brief Bioinform* 6, 3: 239-251.

Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). "A Comparative Evaluation of Term Recognition Algorithms" in *Proceedings of The sixth international conference on Language Resources and Evaluation*, (LREC 2008), Morocco.

[1] See Marín and Rea (2011) for a review on legal English corpora and Rea (2008) on telecommunication corpora.

[2] The *Constitutional Reform Act, 2005* created the Supreme Court which started to work as the court of last resort of the UK in October 2009, until then, it had been the so-called "Law Lords" of the House of Lords who carried out that function. This is the reason why the texts selected from 2008 to 2010 come from both sources.

[3] Drouin's method can be configured so that it also identifies NPs but in this case, this option was deactivated.

[4] Available at: http://termostat.ling.umontreal.ca/index.php?lang=en_CA

[5] Available at: http://terminus.upf.edu

[6] At: http://code.google.com/p/jatetoolkit

[7] We are referring here to those words which do not belong to the specialised domain exclusively like 'abettor' or 'telemetry' but others like 'trial' or 'buzz' which acquire a specialised meaning when in contact with the specific domain, they are the so-called semi-technical words.