Criteria for Measuring Term Recognition

Andy Lauriston
Department of Languages and Linguistics
University of Manchester Institute of Science and Technology
P.O. Box 88
Manchester M60 1QD
United Kingdom
andyl@ccl.umist.ac.uk

Abstract

This paper qualifies what a true term-recognition systems would have to recognize. The exact bracketing of the maximal termform is then proposed as an achieveable goal upon which current system performance should be measured. How recall and precision metrics are best adapted for measuring term recognition is suggested.

1 Introduction

In recent years, the automatic extraction of terms from running text has become a subject of growing interest. Practical applications such as dictionary, lexicon and thesaurus construction and maintenance, automatic indexing and machine translation have fuelled this interest. Given that concerns in automatic term recognition are practical, rather than theoretical, the lack of serious performance measurements in the published literature is surprising.

Accounts of term-recognition systems sometimes consist of a purely descriptive statement of the advantages of a particular approach and make no attempt to measure the pay-off the proposed approach yields (David, 1990). Others produce partial figures without any clear statement of how they are derived (Otman, 1991). One of the best efforts to quantify the performance of a term-recognition system (Smadja, 1993) does so only for one processing stage, leaving unassessed the text-to-output performance of the system.

While most automatic term-recognition systems developed to date have been experimental or in-house ones, a few systems like TermCruncher (Normand, 1993) are now being marketed. Both the developers and users of such systems would benefit greatly by clearly qualifying what each system aims to achieve, and precisely quantifying how closely the system comes to achieving its stated aim.

Before discussing what a term-recognition system should be expected to recognize and how performance in recognition should be measured, two underlying premises should be made clear. Firstly, the automatic system is designed to recognize segments of text that, conventionally, have been manually identified by a terminologist, indexer, lexicographer or other trained individual. Secondly, the performance of automatic term-recognition systems is best measured against human performance for the same task. These premises mean that for any given application - terminological standardization and vocabulary compilation being the focus here - it *is* possible to measure the performance of an automatic term-recognition system, and the best yardstick for doing so is human performance.

Section 2 below draws on the theory of terminology in order to qualify what a true term-recognition system must achieve and what, in the short term, such systems can be expected to achieve. Section 3 specifies how the established ratios used in information retrieval - recall and precision - can best be adapted for measuring the recognition of single- and multi-word noun terms.

2 What is to be Recognized?

Depending upon the meaning given to the expression "term recognition", it can be viewed as either a rather trivial, low-level processing task or one that is impossible to automate. A limited form of term recognition has been achieved using current techniques (Perron, 1991; Bourigault, 1994; Normand, 1993). To appreciate what current limitations are and what would be required to achieve full term recognition, it is useful to draw the distinction between "term" and "termform" on the one hand, and "term recognition" and "term interpretation" on the other.

2.1 Term vs Termform

Particularly in the computing community, there is a tendency to consider "terms" as strictly formal entities. Although usage among terminologists varies, a term is generally accepted as being the "designation of a defined concept in a special language by a linguistic expression" (ISO, 1988). A **term** is hence

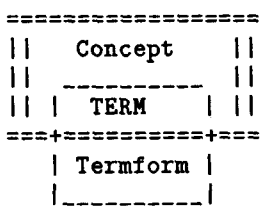

Figure 1: Term vs Termform

the intersection between a conceptual realm (a defined semantic content) and a linguistic realm (an expression or termform) as illustrated in Figure 1. A term, thus conceived, cannot be polysemous although **termforms** can, and often do, have several meanings. As terms precisely defined in information processing, "virus" and "Trojan Horse" are unambiguous; as termforms they have other meanings in medicine and Greek mythology respectively.

This view of a term has one very important consequence when discussing term recognition. Firstly, term recognition cannot be carried out on purely formal grounds. It requires some level of linguistic analysis. Indeed, two term-formation processes do not result in new termforms: **conversion** and **semantic drift**[1]. A third term-formation process, **compression**, can also result in a new meaning being associated with an existing termform[2].

Proper attention to capitalization can generally result in the correct recognition of compressed forms. Part-of-speech tagging is required to detect new terms formed through conversion. This is quite feasible using statistical taggers like those of Garside (1987), Church (1988) or Foster (1991) which achieve performance upwards of 97% on unrestricted text. Terms formed through semantic drift are the wolves in sheep's clothing stealing through terminological pastures. They are well enough concealed to allude at times even the human reader and no automatic term-recognition system has attempted to distinguish such terms, despite the prevalence of polysemy in such fields as the social sciences (Riggs, 1993) and the importance for purposes of terminological standardization that "deviant" usage be tracked. Implementing a system to distinguish new meanings of established termforms would require analyzing discourse-level clues that an author is assigning a new meaning, and possibly require the application of pragmatic knowledge. Until such advanced levels of analysis can be practically implemented, "term recognition" will largely remain "termform recognition" and the failure to detect new terms in old termforms will remain a qualitative shortcoming of all term-recognition systems.

[1]Conversion occurs when a term is formed by a change in grammatical category. Verb-to-noun conversion commonly occurs for commands in programming or word processing (e.g. Undelete works if you catch your mistake quickly). Semantic drift involves a (sometimes subtle) change in meaning without any change in grammatical category (viz. "term" as understood in this paper vs the loose usage of "term" to mean "termform").

[2]Compression is the shortening of (usually complex) termforms to form acronyms or other initialisms. Thus PAD can either designate a resistive loss in an electrical circuit or a "packet assembler-disassembler".

2.2 Term Recognition vs Term Interpretation

The vast majority of terms in published technical dictionaries and terminology standards are nouns. Furthermore, most terms have a complex termform, i.e. they are comprised of more than one word. Sublanguages create series of complex termforms in which complex forms serve as modifiers (*natural language* ⇒ [*natural language*] *processing*) and/or are themselves modified (*applied* [[*natural language*] *processing*]). In special language, complex termforms containing nested termforms, or significant subexpressions (Baudot, 1984), have hundreds of possible syntagmatic structures (Portelance, 1989; Lauriston, 1993). The challenge facing developers of term-recognition systems consists in determining the syntactic and conceptual unity that complex nominals must possess in order to achieve termhood [3].

Another, and it will be argued far more ambitious, undertaking is **term interpretation**. Leonard (1984), Finen (1985) and others have attempted to devise systems that can produce a gloss explicating the semantic relationship that holds between the constituents of complex nominals (e.g. *family estate* ⇒ *estate* **owned by** *a family*). Such attempts at achieving even limited "interpretation" result in large sets of possible relationships but fail to account for all compounds. Furthermore, they have generally been restricted to termforms with two constituents. For complex termforms with three or more constituents, merely identifying how constituents are nested, i.e., between which constituents there exists a semantic relationship, can be difficult to automate (Sparck-Jones, 1985).

In most cases, however, term recognition can be achieved without interpreting the meaning of the term and without analyzing the internal structure of complex termforms. Many term-recognition systems like TERMINO (David, 1990), the noun-phrase detector of LOGOS (Logos, 1987), LEXTER (Bourigault, 1994), etc., nevertheless attempt to recognize nested termforms. Encountering "automatic protection switching equipment", systems adopting this

[3]In this respect, complex termforms, unlike collocations, must designate definable nodes of the conceptual system of an area of specialized human activity. Hence *general trend* may be as strong a collocation as *general election*, and yet only the latter be considered a term.

approach would produce as output several nested termforms (*switching equipment, protection switching, protection switching equipment, automatic protection,automatic protection switching*) as well as the maximal termform *automatic protection switching equipment.* Because such systems list nested termforms in the absence of higher-level analysis, many erroneous "terms" are generated.

It has been argued previously on pragmatic grounds (Lauriston, 1994) that a safer approach is to detect only the **maximal termform**. It could further be said that doing so is theoretically sound. Nesting termforms is a means by which an author achieves transparency. Once nested, however, a termform no longer fulfills the naming function. It serves as a mnemonic device. In different languages, different nested termforms are sometimes selected to perform this mnemonic function (e.g. *on-line credit card checking*, for which a documented French equivalent is *vérification de crédit au point de vente*, literally "point-of-sale credit verification"). Only the maximal termform refers to the designated concept and thus only recognition of the maximal termform constitutes term recognition[4].

Term interpretation may be required, however, to correctly delimit complex termforms combined by means of conjunctions. Consider the following three conjunctive expressions taken from telecommunication texts:

(1) buffer content and packet delay distributions
(2) mean misframe and frame detection times
(3) generalized intersymbol-interference and jitter-free modulated signals

Even the uninitiated reader would probably be inclined to interpret, correctly, that expression (1) is a combination of two complex termforms: *buffer content distribution* and *packet delay distribution.* Syntax or coarse semantics do nothing, however, to prevent an incorrect reading: *buffer content delay distribution* and *buffer packet delay distribution.* Expression (2) consists of words having the same sequence of grammatical categories as expression (1), but in which this second reading is, in fact, correct: *mean misframe detection time* and *mean frame detection time.* Although rather similar to the first two, conjunctive expression (3) is a single term, sometimes designated by the initialism *GIJF.*

Complex termforms appearing in conjunctive expressions may thus require term interpretation for proper term recognition, i.e. reconstructing the conjuncts. If term recognition is to be carried out independently of and prior to term interpretation, as is presently feasible, then it can only be properly seen as "maximal termform recognition" with the meaning of "maximal termform" extended to include the outermost bracketing of structurally ambiguous conjunctive expressions like the three examples above. This extension in meaning is not a matter of theoretical soundness but simply of practical necessity.

In summary, current systems recognize termforms but lack mechanisms to detect new terms resulting from several term-formation processes, particularly semantic drift. Under these circumstances, it is best to admit that "termform recognition" is the currently feasible objective and to measure performance in achieving it. Furthermore, since the nested structures of complex termforms perform a mnemonic rather than a naming function, it is theoretically unsound for an automatic term-recognition system to present them as terms. For purposes of measurement and comparison, "term recognition" should thus be regarded as "maximal termform recognition". Once this goal has been reliably achieved, the output of a term-recognition system could feed a future "term interpreter", that would also be required to recognize terms in ambiguous conjunctive expressions.

3 How Can Recognition be Measured?

Once a consensus has been reached about what is to be recognized, there must be some agreement concerning the way in which performance is to be measured. Fortunately, established performance measurements used in information retrieval - recall and precision - can be adapted quite readily for measuring the term-recognition task. These measures have, in fact, been used previously in measuring term recognition (Smadja, 1993; Bourigault, 1994; Lauriston, 1994). No study, however, adequately discusses how these measurements are applied to term recognition.

3.1 Recall and Precision

Traditionally, performance in document retrieval is measured by means of a few simple ratios (Salton, 1989). These are based on the premise that any given document in a collection is either pertinent or non-pertinent to a particular user's needs. There is no scale of relative pertinence. For a given user query, retrieving a pertinent document constitutes a **hit**, failing to retrieve a pertinent document constitutes a **miss**, and retrieving a non-pertinent document constitutes a **false hit**. **Recall**, the ratio of the number of hits to the number of pertinent documents in the collection, measures the *effectiveness* of retrieval. **Precision**, the ratio of the number of hits to the number of retrieved documents, measures the *efficiency* of retrieval. The complement of recall is **omission** (misses/total pertinent). The complement of precision is **noise** (false hits/total retrieved).

[4]This does not imply that analyzing the internal structure of complex termforms is valueless. It has the very important, but distinct, value of providing clues to paradigmatic relationships between terms.

Ideally, recall and precision would equal 1.0, omission and noise 0.0. Practical document retrieval involves a trade-off between recall and precision.

The performance measurements in document retrieval are quite apparently applicable to term recognition. The basic premise of a pertinent/nonpertinent dichotomy, which prevails in document retrieval, is probably even better justified for terms than for documents. Unlike an evaluation of the pertinence of the content of a document, the term/nonterm distinction is based on a relatively simple and cohesive semantic content[5].User judgements of document pertinence would appear to be much more subjective and difficult to quantify.

If all termforms were simple, i.e. single words, and only simple termforms were recognized, then using document retrieval measurements would be perfectly straightforward. A manually bracketed term would give rise to a hit or a miss and an automatically recognized word would be a hit or a false hit. Since complex termforms are prevalent in sublanguage texts, however, further clarification is necessary. In particular, "hit" has to be defined more precisely. Consider the following sentence:

The latest committee draft reports progress toward constitutional reform.

A terminologist would probably recognize two terms in this sentence: *committee draft* and *constitutional reform*. The termform of each is complex. Regardless of whether symbolic or statistical techniques are used, "hits" of debatable usefulness are apt to be produced by automatic term-recognition systems. A syntactically based system might have particular difficulty with the three consecutive cases of noun-verb ambiguity *draft*, *reports*, *progress*. A statistically based system might detect *draft reports*, since this cooccurrence might well be frequent as a termform elsewhere in the text. Consequently, the definition of "hit" needs further qualification.

3.2 Perfect and Imperfect Recognition

Two types of hits must be distinguished. A **perfect hit** occurs when the boundaries assigned by the term-recognition system coincide with those of a term's maximal termform ([*committee draft*] and [*constitutional reform*] above). An **imperfect hit** occurs when the boundaries assigned do not coincide with those of a term's maximal termform but contain at least one wordform belonging to a term's maximal termform. A hit is imperfect if bracketing either includes spurious wordforms ([*latest committee draft*] or [*committee draft reports*]), fails to bracket a term constituent (*committee* [*draft*]) or both (*committee* [*draft reports*]). Bracketing a segment containing no wordform that is part of a term's maximal termform is, of course, a false hit ([*reports progress*]).

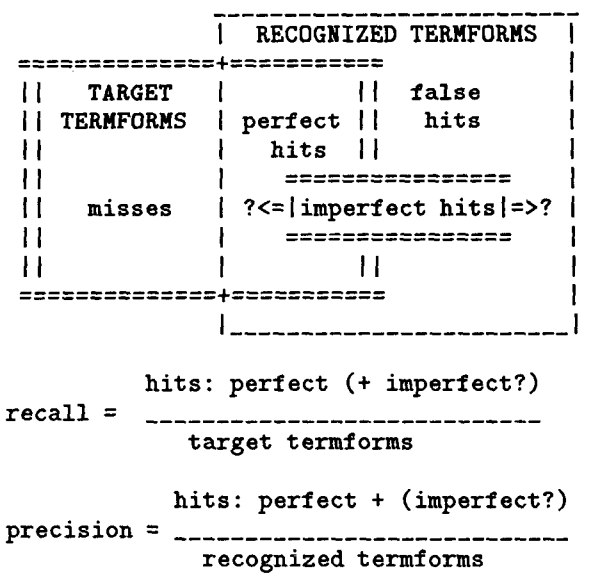

Figure 2: Recall, Precision and Imperfect Hits

The problematic case is clearly that of an imperfect hit. In calculating recall and precision, should imperfect hits be grouped with perfect hits, counted as misses, or somehow accounted for separately (Figure 2)? How do the perfect recall and precision ratios compare with imperfect recall and precision (including imperfect hits in the numerator) when these performance measurements are applied to real texts? Counting imperfectly recognized termforms as hits will obviously lead to higher ratios for recall and precision, but how much higher?

To answer these questions, a complex-termform recognition algorithm based on weighted syntactic term-formation rules, the details of which are given in Lauriston (1993), was applied to a tagged 2,861 word text. The weightings were based on the analysis of a 117,000 word corpus containing 11,614 complex termforms as determined by manual bracketing. The recognition algorithm includes the possibility of weighting of the terminological strength of particular adjectives. This was carried out to produce the results shown in Figure 3.

Recall and precision, both perfect and imperfect, were plotted as the algorithm's term-recognition threshold was varied. By choosing a higher threshold, only syntactically stronger links between adjacent words are considered "terminological links". Thus the higher the threshold, the shorter the average complex termform, as weaker modifiers are

[5]In practice, terminologists have some difficulty agreeing on the exact delimitation of complex termforms. Still five experienced terminologists scanning a 2,861 word text were found to agree on the identity and boundaries of complex termforms three-quarters of the time (Lauriston, 1993).

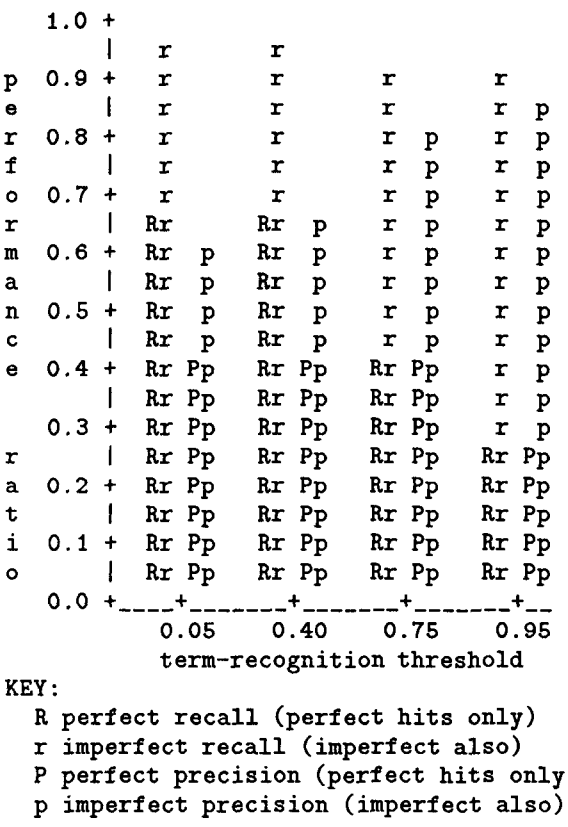

Figure 3: Effect of Imperfect Hits of Performance Ratios

stripped from the nucleus. Lower recall and higher precision can be expected as the threshold rises since only constituents that are surer bets are included in the maximal termform.

This Figure 3 shows that both recall and precision scores are considerably higher when imperfect hits are included in calculating the ratios. As expected, raising the threshold results in lower recall regardless of whether the ratios are calculated for perfect or imperfect recognition. There is a marked reduction in perfect recall, however, and only a marginal reduction in imperfect recall. The precision ratios provide the most interesting point of comparison. As the threshold is raised, imperfect precision increases just as the principle of recall-precision tradeoff in document retrieval would lead one to expect. Perfect precision, on the other hand, actually declines slightly. The difference between perfect and imperfect precision (between the P-bar and p-bar in each group) increases appreciably as the threshold is raised. This difference is due to the greater number of recognized complex termforms either containing spurious words or only part of the maximal termform.

Two conclusions can be drawn from Figure 3. Firstly, the recognition algorithm implemented is poor at perfect recognition (perfect recall $\approx$ 0.70; perfect precision $\approx$ 0.40) and only becomes poorer as more stringent rule-weighting is applied. Secondly, and more importantly for the purpose of this paper, Figure 3 shows that allowing for imperfect bracketing in term recognition makes it possible to obtain artificially high performance ratios for both recall and precision. Output that recognizes almost all terms but includes spurious words in complex termforms or falls short of recognizing the entire termform leaves a burdensome filtering task for the human user and is next to useless if the "user" is another level of automatic text processing. Only the exact bracketing of the maximal termform provides a useful standard for measuring and comparing the performance of term-recognition systems.

4 Conclusion

The term-recognition criteria proposed above - measuring recall and precision for the exact bracketing of maximal termforms - provide a basic minimum of information needed to assess system performance. For some applications, it is useful to further specify how these performance ratios differ for the recognition of simple and complex termforms, how they vary for terms resulting from different term-formation processes, what the ratios are for termform types as opposed to tokens, or how well the system recognizes novel termforms not already in a system lexicon or previously encountered in a training corpus. Precision measurements might usefully state to what extent errors are due to **syntactic noise** (bracketing crossing syntactic constituents) as distinguished from **terminological noise** (bracketing including nonclassificatory modifiers or omitting classificatory ones).

Publishing such performance results for term-recognition systems would not only display their strengths but also expose their weaknesses. Doing so would ultimately benefit researchers, developers and users of term-recognition systems.

References

Baudot, Jean, and André Clas (1984). A model for a bilingual terminology mini-bank. In *Lebende Sprachen*, Vol.19, No.2, pp.283–298.

Black, Ezra, Roger Garside, and Geoffrey Leech (1993). *Statistically-driven computer grammars of English: the IBM/Lancaster approach.* Amsterdam: Rodopi.

Bourigault, Didier (1992). LEXTER, un Logiciel d'EXtraction de TERminologie. In *Colloque sur le repérage de l'information textuelle*, Montréal, Québec, Hydro-Québec, March, pp.15–25.

Bourigault, Didier (1994). Extraction et structuration automatique de terminologie pour l'aide à l'acquisition des connaissances à partir de textes. In *Reconnaissance des formes et intelligence artificielle (RFIA'94)*, Paris, January.

Church, Kenneth W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas. Association for Computational Linguistics, Morristown, New Jersey, pp.136–143.

Daille, Béatrice (1994). *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques.* Université de Paris 7 (PhD Thesis), Paris, January.

David, Sophie and Pierre Plante (1990). De la nécessité d'une approche morphosyntaxique en analyse de textes. In *Intelligence artificielle et sciences cognitives au Québec*, Vol.3 No.3, September, pp.140–155.

Finin, Timothy W. (1986). Constraining the interpretation of nominal compounds in a limited context. In Ralph Grishman and Richard Kittredge, editors, *Analyzing language in restricted domains*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, pp.163–173.

Foster, George F. (1991). *Statistical lexical disambiguation.* McGill University (MSc Thesis), Montréal, Québec.

Garside, Roger, Geoffrey Leech, and Geoffrey Sampson, editors (1987). *The computational analysis of English: a corpus-based approach*, London, Longman.

Gaussier, Éric and Jean-Marc Langé (1994). Some methods for the extraction of bilingual terminology. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, Manchester, England, University of Manchester Institute of Science and Technology (UMIST), September, pp.242–247.

International Organization for Standardization (ISO) (1988). *Terminolgy - vocabulary.* (ISO/DIS 1087), Geneva, Switzerland.

Jones, Leslie, Edward W. Gassie, and Sridhar Radhakrishnon (1990). INDEX: the statistical basis for an automatic conceptual phrase-indexing system. In *Journal of the American Society for Information Science*, Vol.41, No.2, pp.87–97.

Lauriston, Andy (1993). *Le repérage automatique des syntagmes terminologiques.* Université du Québec à Montréal (MA Thesis), Montréal, Québec.

Lauriston, Andy (1994). Automatic recognition of complex terms: problems and the "Termino" solution. In *Terminology: applications in interdisciplinary communication*, Vol.1, No.1, pp.147–170.

Leonard, Rosemary (1984). *The interpretation of English noun sequences on the computer*, Amsterdam, North-Holland.

Logos Corporation (1987). *LOGOS English Source Release 7.0*, Dedham, Mass., Logos Corporation.

Normand, Diane (1993). Quand la terminologie s'automatise: du nouveau en terminotique: Term Cruncher, un logicel de dépouillement. In *Circuit*, No.42, December, pp.29–30.

Otman, Gabriel (1991). Des ambitions et des performances d'un système de dépouillement terminologique assisté par ordinateur. In *La Banque des mots* (special issue on terminology software), No.4, pp.59–96.

Perron, Jean (1991). Présentation du progiciel de dépouillement terminologique assisté par ordinateur: Termino. In *Les industries de la langue: perspectives des années 1990*, Montéal, Québec, Office de la langue française/Société des traducteurs du Québec, pp.715–755.

Portelance, Christine (1989). *Les formations syntagmatiques en langues de spécialité.* Université de Montréal, (PhD Thesis), Montréal, Québec.

Riggs, Fred (1993). Social science terminology: basic problems and proposed solutions. In Helmi B. Sonneveld and Kurt T. Loening, editors, *Terminology: applications in interdisciplinary communication*, Amsterdam, John Benjamins, pp.195–222.

Salton, Gerard (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*, Reading, Mass., Addison-Wesley.

Smadja, Frank (1993). Retrieving collocations from text: Xtract. In *Computational linguistics*, Vol.19, No.1.

Spark-Jones, Karen (1985). Compound Noun Interpretation Problems. In Frank Fallside and William A. Woods, editors, *Computer Speech Processing*, Englewood Cliffs, New Jersey, Prentice-Hall, pp.363–381.

Van der Eijk, Pik (1993). Automating the acquisition of bilingual terminology. *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pp.113–119.