KB-N: Automatic Term Extraction from Knowledge Bank of Economics

Magnar Brekke
Kai Innselset
Marita Kristiansen
Kari Øvsthus

Center for LSP Research/Dept. of Professional and Intercultural Communication/NHH
Helleveien 30, 5045 Bergen, Norway
Magnar.Brekke/Kai.Innselset/Marita.Kristiansen/Kari.Ovsthus@nhh.no

Abstract

KB-N is a web-accessible searchable Knowledge Bank comprising A) a parallel corpus of quality assured and calibrated English and Norwegian text drawn from economic-administrative knowledge domains, and B) a domain-focused database representing that knowledge universe in terms of defined concepts and their respective bilingual terminological entries. A central mechanism in connecting A and B is an algorithm for the automatic extraction of term candidates from aligned translation pairs on the basis of linguistic, lexical and statistical filtering (first ever for Norwegian). The system is designed and programmed by Paul Meurer at Aksis (UiB). An important pilot application of the term base is subdomain and collocations based word-sense disambiguation for LOGON, a system for Norwegian-to-English MT currently being developed.

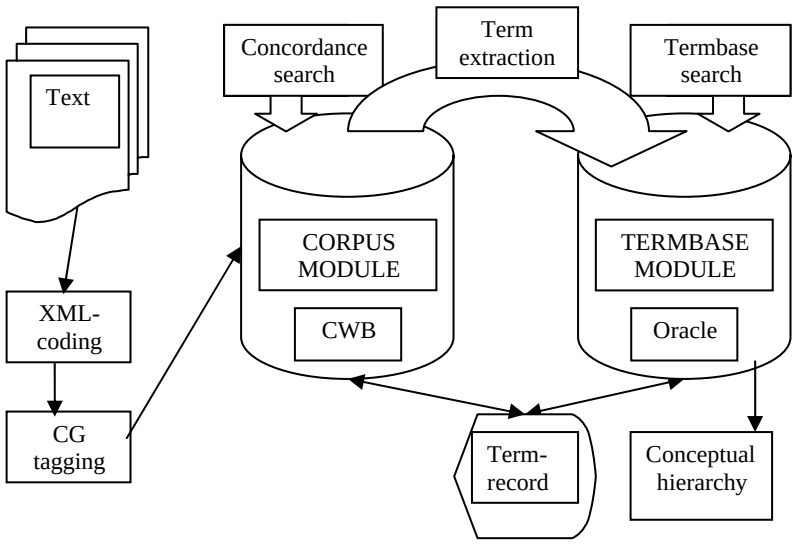

Figure 1: KB-N System architecture

1. Foundations

KB-N (KnowledgeBank of Norway) proceeds from the assumption that specialist knowledge of a given domain/subdomain resides in text, which allows that knowledge to be fixated, managed and conveyed. We further assume that the gateways to such subdomain knowledge are made up of concepts which have definable links to related concepts, together making up conceptual structures. In principle such a concept can be seen as a gate which remains closed for non-specialists until they discover or are taught the appropriate term (in a given language) which can unlock the gate and access the essential meaning. Conversely, specialist researchers having just made a discovery or completed a theoretical analysis may be unable to communicate their results until a suitable term is created (in a given language).

It follows from these assumptions that by capturing professional text produced by specialists of a given subdomain we may be reasonably sure that we capture an essential subset of their subdomain knowledge – hence the motivation for establishing balanced corpora of subdomain texts. By identifying the terminological access points in such a text, by further linking each term to its concept and providing a standard description of its meaning (i.e. definition) we gain a platform for managing essential terminological information (i.e. variants, contexts, collocations, equivalents in other languages etc.) – hence the motivation for establishing term records and arranging them in searchable termbanks.

It should be emphasized that the subdomain corpus and the termbank manifesting its essential knowledge are complementary and should remain a dynamically linked specialist language resource. In its fundamental conception the KB-N database bears considerable affinity to the thinking which seems to underlie the GENOMA-KB (cf. Cabre et al. 2004).

Finally it should also be emphasized that in placing KB-N squarely in the context of LSP, i.e. Language for Special Purposes (aka *Fachsprache*), rather than that of

LGP or general language as such, we can allow ourselves to focus entirely on the specific aspects characterizing subdomain languages (cf. Brekke 2004, p. 41).

2. Implementations

In the language resource arena KB-N is constructing the two major components just indicated. The overall architecture will appear as represented in Figure 1.

2.1. Subdomain Corpus Module

One major module consists of a comprehensive English/Norwegian corpus of professional text representing relevant document types and text genres of economic-administrative knowledge domains. Of primary interest are expository, didactic, popularized and regulatory texts reflecting different levels of expertise. The ultimate text volume is envisaged at about 30m words (each language 15m) from various authoritative sources.

Initial focus has been on finding parallel texts in English and Norwegian where one is a certified translation of the other. Using University of Stuttgart's Corpus WorkBench[1] (and Oracle as a platform) each text has been XML-coded and word-class-tagged (using the Oslo-Bergen Tagger[2]). The parallel text versions have then been aligned via Hofland's lexical anchor method[3] and made available for routine as well as user-initiated bilingual parallel concordancing, an essential feature of automatic term extraction (to be described below).

Increasingly, however, the text base will include monolingual texts representing identical subdomains and communication types, given that the supply of strictly parallel texts is rather limited when Norwegian is the other member of a language pair. Consideration of text for inclusion in the corpus requires careful scrutiny of stylistic quality, lexical representativity as well as conceptual substance, and, in the case of parallel texts, the professional quality and equivalence of the target text must be assessed.

2.2. Termbase Module

The other major component is our concept-oriented bilingual terminological database, a repository of term records (entries) of domain-specific knowledge extracts from the corpus. The emergence of computerized corpus-based methods has of course had an enormous impact on terminology research but without entirely displacing the time-honored technique of excerpting by hand. In fact the general problems of "silence" and "noise" in terminology extraction (well described in Castellvi et al. (2001)) invite us to view the two approaches as complementary, by allowing relevant items not represented in the text samples to be supplied by a subdomain expert.

The term record accommodates domain specific term equivalents, synonyms, acronyms etc. in the respective languages and links them to their common concept. For each central concept a standard definition is provided and its relative position in the concept structure (whether hierarchic, cognitive or otherwise) is indicated. The pivotal role here played by the concept facilitates future inclusion of other languages in the term bank. The conceptual structures themselves appear in a separate window where they can be established, inspected and manipulated.

For each term one or more characteristic authentic usage contexts are given, and to aid (automatic) word sense disambiguation (essential for MT) a set of domain-specific collocations are listed. The link between term and concept must be established by a domain expert, whose tacit knowledge is also required for the identification of "missing" concepts based on the systematization of conceptual structures.

Other than such input the remaining knowledge represented in the term record is either extracted from or in the main based on the corpus text samples. The basic mechanism involved is the automatic extraction of term candidates, to which we now turn.

3. Term Extraction

While automatic term extraction (ATE) from English is beginning to be well researched from various theoretical and computational angles, most of the specific techniques proposed and tested (but especially linguistic ones) are strongly sensitive to typological differences between languages. Thus the strategies available for English differ markedly from those relevant for Romance languages, or for those of stricter Germanic stock, which is the case for Norwegian. We have not come across published work indicating that Norwegian ATE has been tackled before; see Øvsthus (2005). Our approach is three-pronged, exploiting linguistic, lexical, as well as statistical techniques, see Table 1.

1. linguistic filter:
a) regular expressions
(adj. in positive form)* + noun [minus genitive form] adjective + "*og/eller*" + adjective + noun noun + "-" + "" + noun noun + "*og/eller*" + "-"noun
b) general vocabulary trap
2. Named Entity Recognizer:
Salvages strings having failed the linguistic filter according to specific criteria
3. Statistical Significance ("Weirdness") ratio
Text occurrence ratio checked against occurrence ratio in major GL corpus .

Table 1: Norwegian Term Candidate Extraction

3.1. Linguistic Filters

ATE from a given Norwegian text starts from a fairly straightforward identification of complex noun phrases (CNPs), on the general assumption that the

[1] http://www.ims.uni-stuttgart.de
[2] See Hagen et al. 2000
[3] See Hofland 1996

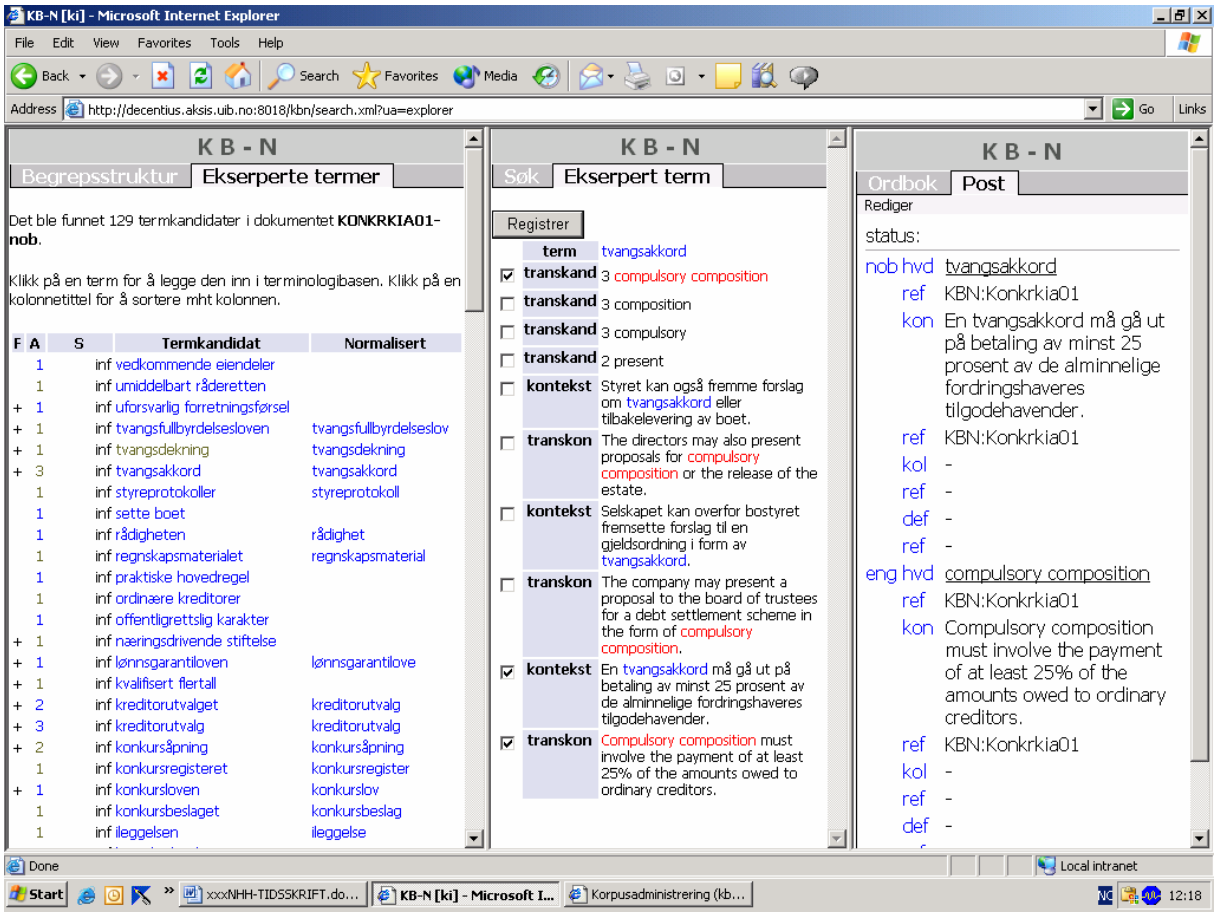

Figure 2: KB-N Termbase window; term candidates (left), selection template (middle) and term record (right)

overwhelming majority of technical terms are nominal (an empirically well attested observation in many languages). CNPs can be extremely complex, especially in English, but again the Germanic typology of Norwegian gives us an advantage: Concatenation of nominal roots can result in very long words, as is often the case in LSP word formation, but our analysis need not go beyond the identification of inflectional morphology.

As indicated in Table 1 admissible Norwegian complex NPs will have an uninflected noun with optional preposed adjective(s) in the positive form. Adjectives as well as nouns can be conjoined, and hyphenated expressions saved for later splitting (e.g. *valutagevinst og -tap* is given two entries: *valutagevinst*, *valutatap*.

The lexical trap is a cumulative stop-list of non-focal adjectives like *adskillige* ("several"), *foreløpig* ("preliminary") and *øvrige* ("further"), i.e. very general modifiers devoid of domain-specific content.

The Named Entity Recognizer has been inherited from an independent project called *Nomen Nescio*,[4] a fairly standard NER-algorithm which salvages sequences of e.g. institutional names containing a mixture of content words and function words which would otherwise fail to be picked up.

3.2. Statistical Filter

It is a well established fact that terms in running text often display conspicuous behavior, either because of a) high frequency or b) "weirdness",[5] i.e. low frequency where occurrence at all is unusual or unexpected. First an occurrence ratio is calculated for all strings having survived the linguistic filtering of a given LSP text. As a standard of comparison one needs access to the occurrence ratios of all words in a large volume of LGP text

In our case Hofland's cumulative corpus of general Norwegian newspaper text (currently near 400m words; see Hofland n.d.) is accessed and compared with the ratios generated from the new text, and a salience ratio is calculated for items exceeding a set threshold level. The end result of this filtering process is a list of "recommended" term candidates presented to the human expert for confirmation/rejection before final inclusion into the term bank.

Figure 2 is a snapshot of the KB-N Termbase window. The leftmost frame is displaying the top 23 (of 129) term candidates proposed by our ATE algorithms, here sorted according to salience ratio. The label "inf" (=infinite) under the S-column reflects the fact that the item in the "Termkandidat"-column does not appear in Hofland's reference corpus of general Norwegian; items lower on the list will display figures like "19624.95" or "87.22", which indicate an item which appears there with fairly low or fairly high frequency, respectively.

[4] http://scrooge.spraakdata.gu.se/nn/

[5] Ahmad & Rogers 2001

The term candidate list can also be sorted alphabetically or according to frequency (under column A). "+" in the leftmost column shows that the item already has an entry in the KB-N Termbank.

The middle frame in Figure 2 appears after the item *tvangsakkord* has been pressed in the candidate list. This particular selection template has been ticked of for "Transkand" *compulsory composition* (automatically suggested translation English equivalent for *tvangsakkord)* along with a suitable context for each language, all of which extracted automatically from the parallel concordance KB-N produces on the basis of its strictly aligned parallel text. When the "Registrer"-button is pressed, the selected material will turn up directly in the corresponding term record (shown in the rightmost frame), ready to be stored as a new entry. To achieve this the human operator has pressed the left mouse button a total of six times in the course of a few seconds.

The volume of term records currently held in the KB-N KnowledgeBank is about 5000, a situation attributable not only to the continuous development and refinement of automatic term extraction, but in large measure also to the efficiency of the tools for human-machine interaction just described, which have been designed to optimize the work flow in the KB-N project. We have consciously avoided making the term selection and entry procedure fully automatic, to avoid inundating the termbase with noise and junk which eventually would require considerable effort to eliminate.

4. Hook-up with MT

LOGON[6] is a massive effort to develop MT from Norwegian to English which, when completed, would stand to benefit from having access to the subdomain knowledge residing in the KB-N Termbase. While the sheer complexity of systems makes it unlikely that there will be a direct link between the two, plans are under way for entering KB-N terms along with subdomain and collocations information into the LOGON lexicons for analysis, transfer and generation. Subdomain and collocations information shows considerable promise in partly alleviating the perennial achilles' heel of MT, namely Word Sense Disambiguation, when translating subdomain LSP text rather than LGP (see Magnini et al. 2005, McRoy 1992, Yarowsky 1993).

5. Conclusion

We consider automatic term extraction the computationally most interesting achievement of the KB-N project so far, in exploiting the empirical value of linguistic resources (acquired for quite different purposes) in developing precise algorithms for automatically generated term candidate lists. This operation constitutes a significant link between the text bank and the term bank and exploits human-machine interaction to combine text-embedded domain knowledge with human expertise in a form which can be utilized in e.g. MT, e-learning, human translation, and knowledge management.

[6] http://www.emmtee.net/

6. References:

Ahmad K. & M. Rogers (2001). Corpus linguistics and terminology extraction. In Ahmad, Wright, Rogers & Budin (eds). *Handbook of terminology management*, vol. 2. Philadelphia: Benjamins; p. 725-60.

Brekke, M. (2004) "KB-N: Computerized extraction, representation and dissemination of special terminology". Costa, R. et al. (eds). *Workshop on Computational and Computer-assisted Terminology, IV LREC 2004*, Lisbon.

Cabré, M.T. et al. (2004). The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities. Lisbon: *IV LREC 2004*. p. 87-90.

Castellvi M.T.C., Bagot R.E & J. Vivaldi Palatresi (2001). Automatic term detection: a review of current systems. In: Bourigault D. et al. (eds). *Recent advances in computational terminology*. Philadelphia: Benjamins;. p. 53-87.

Hagen K., J. Bondi Johannessen & A. Nøklestad (2000). "A Constraint-Based Tagger for Norwegian". In Lindberg, C.-E. og S. Nordahl Lund (eds). *17th Scandinavian Conference of Linguistics*, *Odense Working Papers in Language and Communication*, No. 19, vol I.

Hofland, K. (1996) A program for aligning English and Norwegian sentences. In S. Hockey et al. (eds), *Research in Humanities Computing*. Oxford: OUP. 165-178.

Hofland, K. (n.d.) Norwegian Newspaper Corpus at http://helmer.aksis.uib.no/aviskorpus/.

Ide, N. & J. Veronis (1998). Word Sense Disambiguation: the state of the art. *Computational Linguistics*, 24(1), pp. 1-41.

Kristiansen, M. (2005). Disciplinary autonomy and concept relations in electronic knowledge bases. A theoretical approach to KB-N - a knowledge base for economic-administrative domains. *SYNAPS - Fagspråk. Kommunikasjon. Kulturkunnskap,* 17(2005). Bergen: NHH, pp.1-7.

Magnini, B. et al. (2005). The role of domain information in word sense disambiguation. *Natural Language Engineering* 8 (1), pp. 1.14.

McRoy, S.W. (1992). Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics* 18(1), pp. 1.30.

Yarowsky, D. (1993). One sense per collocation. *Proceedings of 5th DARPA Speech and Natural Language Workshop.*

Øvsthus, K. et al. (2005). Developing automatic term extraction. Automatic domain specific term extraction for Norwegian. *Proceedings of Terminology and Knowledge Engineering*, Copenhagen: CBS.