John Laffling
University of Wolverhampton

An Analogical Dictionary for Machine Translation*

Abstract

Memory and analogy are beginning to dethrone calculative rationality in semi– and fully automated translation systems. This paper provides an outline of the design, construction and implementation of the lexicographical component of one such system, which, used in conjunction with a database of translation examples, aligned at the sentential and sub– sentential level, enables the computer to exploit analogical reasoning procedures to translate new input for which no literal match can be found. Comparisons with the lexicographical components of other systems within the same paradigm are made.

1. Introduction

This paper discusses the design, organization and application of the lexicographical component of a German–English machine translation system, which is currently under construction and is designed to translate texts from the domain of environmental issues. This component takes the form of a thesaurus–like dictionary, which, used in conjunction with a database composed of thousands of source language examples (some 4,000[1] at present), at the level of the sentence, clause and syntagma, together with their target language correspondents, enables the computer to reason by analogy to translate new input for which no literal match can be found. The system is written in PDC Prolog, and is currently running on an Amstrad PC7486 with 8MB RAM.

2. Design and organization

In the Introduction, the dictionary was not referred to explicitly as a thesaurus, for the simple reason that it is only in terms of content, but not organization, that it warrants this appellation. The present dictionary is not confined to hierarchical relationships, but, motivated by a desire to tap a source of power at the heart of human intelligence, i.e. the ability to reason by analogy, attempts to *classify* lexical items (see Section 3.2) in a more complex manner, going beyond the conventional relations of synonymy, hyponymy and meronymy, to include, for example, antonymy (both direct and indirect), indexal relations (e.g. Ozonabbau: Klimaveränderungen) and transformational relations (e.g. Zerstörung: zerstören; verpflichtet sein: die Verpflichtung haben). Exclusively hierarchical organization was deemed to entail creation of arbitrary categories, leading, in some cases, to assignment of lexical items to categories into which they fit only uneasily. In the *Longman Lexicon of Contemporary English*, for example, an on–line version of which is used by Sato (1991b) and Sumita et al. (1993), one finds *disband* seemingly forced, for lack of a more appropriate head, under which it could be linked with, say, *send home, break up, demobilize, split up and unite*, into a grouping that contains *dismiss, discharge, sack*, and *fire*, under the category DISMISSING AND RETIRING PEOPLE (see also Michiels and Noël (1982)). Furthermore, in this same lexicon at least, different senses of words are stored under the same category (e.g. *animate*, in its second sense of 'make exciting', is located under the category EXISTING AND CAUSING TO EXIST).

Rather than imposing a hierarchical structure[2] on the dictionary entries, therefore, a decision in favour of alphabetical arrangement was taken, only such lexical information as was corroborated by corpus–based evidence being assigned to entries. To this extent, the present dictionary follows the proposal advocated by Agricola (1972:14):

> 'Das Attribut 'sprachlicher Thesaurus' soll einerseits betonen, daß die Sememe nicht nach extralingualen Gesichtspunkten in eine logisch–begriffliche Ordnung zu arrangieren oder einem anderen natürlichen bzw. vorweggenommenen System der Realität zuzuordnen sind, sondern daß deren Widerspiegelung in Form *innersprachlich objektiv feststellbarer Gegebenheiten* als Grundlage dienen.' [Emphasis added].

The redundancy typically associated with alphabetical ordering is avoided, to a large extent, by utilization of a pointer system, which, in obviating the need for information stored under one headword to be explicitly repeated under another, allows a much more compact and efficient organization of the lexical data. The following example may serve as an illustration of the lexical information coded for nouns, adjectives[3] and verbs.

Example 1

Zunahme (n)
$ Steigerung $ Erhöhung $ Anstieg * Abnahme * Verringerung
> Temperaturzunahme
schädlich (adj)
$ schädigend * unschädlich * harmlos * ungefährlich
< nachteilig > umweltschädlich > ozonschädlich
> gesundheitsschädlich > klimaschädlich
festlegen (v;sep)
# Festlegung $ festsetzen $ bestimmen > festschreiben
[# = transformational; $ = synonym;[4] < = hyperonym; > = hyponym;* = antonym]

The headword thus holds a list of related words, which are, in turn, linked to this set by means of pointers.

3. Related work

The present research is to be seen in the light of the recent increasing interest accorded, by a (small) number of researchers in the machine translation community (e.g. Sadler (1989), Sato (1991a; 1991b; 1993), Sato and Nagao (1990), Sumita et al. (1991; 1993), Furuse and Iida (1992)), to example–based translation, an approach in which previously translated text fragments are viewed as examples on which to base the translation of new input. Whilst the work cited has been foundational to the approach adopted here, there are, however, significant differences.

3.1 Domain–specific vs general

Where the above projects have access to external knowledge structures (Sadler (1989) is alone in foregoing such a structure, proposing to keep meaning entirely *implicit* by representing it in terms of the contextual overlap between syntactically and referentially aligned bilingual texts), they take the form of hand–crafted, already existing word hierarchies of the general language (in machine–readable form); the present lexicon, on the other hand, is domain–specific (relating to party political texts on environmental issues) and has been (and is being[5]) compiled semi–automatically from a large corpus of texts (currently more than 2 million words). Procedures for partial automation include, for example, provision, via a menu–driven interface to concordance facilities, for allowing the lexicographer–cum–user to enter, into the dictionary, any of the expressions appearing in the surrounding context of a particular searchword; s/he is asked under which entry s/he would like them to be incorporated, and by which symbol they should be prefaced.

Given the types of text that MT systems are required to translate, the EBMT paradigm will, if it is to make significant advances, soon have to confront the task of constructing domain–specific thesauruses.[6] Nor is it the case, as is sometimes tacitly assumed (see Sato (1991a:4), that domain–specific knowledge can simply be integrated into general thesauruses: this additional information (that, in the present corpus, for example, *abbaubar* and *persistent* are antonyms (as corroborated by 'der Eintrag von *persistenten* Stoffen, also im Boden *nicht abbaubaren* problematischen Stoffen') and that *Weltbank* is related hyponymously to *Geber* (as evidenced by 'die *Weltbank* und andere *Geber* unterstützen die Ausarbeitung nationaler Umwelt– Aktionspläne) may well be at odds with already existing data.

3.2 Explicit naming vs grouping

As will have become evident from Section 2 above, relationships are explicitly named in the present dictionary. As, however, existing thesauruses simply group related words without attempting to classify each relationship, EBMT systems that utilize them exploit the codes assigned to categories to calculate similarities between words. As Furuse and Iida (1992:97) state:

> The distance between two words is reduced to the distance between their respective semantic attributes in a thesaurus. Words have associated thesaurus codes, which correspond to particular semantic attributes. The distance between the semantic attributes is determined according to the relationship of their positions in the hierarchy of the thesaurus[7], and varies from 0 to 1.

The difference in the structuring of semantic knowledge leads to different execution of similarity matches. The present system does not require an explicit similarity metric, the customary (and possibly expensive) calculations carried out by other systems being replaced by provision of knowledge of weightings assigned to the named relations, as follows (in order of prioritization): identity, transformationals, synonymy, hyperonymy/ hyponymy co–hyponymy, antonymy, meronymy/holonymy, indexal etc.[8]

3.3 Phrasal units and transformational relations

The present dictionary also differs from those used in other systems in that it includes a) phrasal units[9] (e.g. *stoffliche Verwertung* is coded as being co–hyponymous with *thermische Verwertung*; *gefährliche Stoffe* as being synonymous with *Gefahrstoffe*; and *gentechnische Methoden* as being synonymous with *Methoden der Gentechnik*), giving rise to improved matching performance, and b), as already mentioned in Section 1, semantic equivalences that obtain between elements of varying syntactic function, the so–called transformationally related items. Transformational relations (e.g. *gentechnisch veränderte Organismen sollen nur freigesetzt werden: die Freisetzung gentechnisch veränderter Organismen*) prove to be one of the most frequently employed means of reiteration in the corpus.

4. Application

In this Section, I will illustrate the application of the dictionary component, outlining its role both in determination of the closest semantic analogue and in selection of appropriate translation correspondents, by means of the following input (taken from a text which was not used for the purposes of deriving database examples): *Der Kraftstoffverbrauch muß durch Festlegung von Grenzwerten verringert werden.*

Once a global structure (i.e. NP modal PP pastpart auxinf) has been assigned to this new input, the first step is to call up examples in the database which are isomorphic[10] with it. The result,[11] in the present case, is as follows (English correspondents have been omitted, owing to considerations of space):

Example 2

a) Diese Probleme können durch gemeinsame Anstrengungen aller Staaten gelöst werden;

b) International abgestimmte Regelungen sollen durch nationale Maßnahmen ergänzt werden;

c) Die Beschaffenheit von Stoffen und Erzeugnissen kann auf der Grundlage des Gesetzes geregelt werden;

d) Die wirklichen Kosten des Wirtschaftens müssen bei den Verursachern berücksichtigt werden;

e) Die Schadstoffbelastung durch Kraftfahrzeuge muß durch fortdauernde Konstruktionsverbesserungen reduziert werden[12]

Next, the closest semantic analogue is selected; nouns and verbs (the main information-bearing units) in the new input are clamped and, together with their semantic correlates, used as a probe into the list of examples previously called up. In this case, given that the dictionary contains information to the effect that *verringern* and *reduzieren* are synonymous, the final example in the above list is chosen. This example is coded with the following morpho-syntactic information and alignment possibilities.[13]

Example 3

1 Die (defart;sg;nom)
2 Schadstoffbelastung (n;sg;nom)
3 durch (p)
4 Kraftfahrzeuge (n;pl;acc)
5 muß (modal;sg;3rd)
6 durch (p)
7 fortdauernde (adj;pl;acc)
8 Konstruktionsverbesserungen (n;pl;acc)
9 reduziert (pastpart)
10 werden (aux;inf)

1 motor (n)
2 vehicle (n;sg)
3 pollution (n;sg)
4 must (modal)
5 be (aux;inf)
6 reduced (pastpart)
7 by (p)
8 continued (adj)
9 improvements (n;pl)
10 in (p)
11 construction (n;sg)

1–4=1–3; 2–4=1–3; 5,9,10=4–6; 6–8=7–11; 7–8=8–11; 8=9–11

This closest analogue serves, in turn, as a template, providing a target pattern and, by showing which constituents can be substituted, guiding the next stage of the process, i.e. chunking and recombination.[14] As the database contains no other TL equivalent for *verringert*, which has been found to be synonymous with *reduziert*, the translation unit *muß reduziert werden* is adopted wholesale. Next, searches are carried out for instances of

Kraftstoffverbrauch (and semantic correlates, as suggested by the dictionary, (e.g. *Treibstoffverbrauch, Benzinverbrauch*),[15] the system returning 'fuel consumption', on the basis of 'der spezifische *Kraftstoffverbrauch* eines Autos sagt nichts über die Höhe des Schadstoffausstosses des Motors aus' ('a car's specific *fuel consumption* ...').

A prepositional phrase sequence containing *Festlegung* (or semantic correlates (i.e. $ Bestimmung $ Festsetzung > Festschreibung)) is then used as a probe into the example database, producing the following closest analogue (reproduced, this time, without coding or numbering and with only the essential alignments):

Example 4

> Die SPD fordert eine Wende in der europäischen Verkehrspolitik *durch die Festlegung von ökologischen Mindeststandards.* (The SPD calls for a change in European transport policy *by setting ecological minimum standards*)
>
> Alignments: 10–15=11–15; 10,12,13,15=11,12,14,15; 15=14,15; 12=12

Now that the target pattern for the global unit has been identified, correspondents for the translatable units *Festlegung* and *Grenzwert* are sought, with a view to inserting them into their respective place holders, as indicated by the above alignment possibilities. As Example 5 illustrates, however, the results of the search confront the system with an additional problem, in the case of *Festlegung*: that of selection between several TL correlates.

Example 5

a) Die Festlegung von Eckwerten für den Kraftstoffverbrauch
(the specification of benchmark figures for fuel consumption)

b) Festlegung von Standorten für Kernkraftwerke
(determination of nuclear power station sites)

c) Festlegung von Verbrauchsobergrenzen der Kraftfahrzeuge
(establishment of upper limits on consumption by motor vehicles)

The analogical dictionary is, therefore, called upon to assess the degree of semantic proximity between *Grenzwert* and each of the collocational partners of *Festlegung* found in the examples (i.e. *Eckwert, Standort, Verbrauchsobergrenze, Mindeststandard*).[16].

Eckwert is selected on the basis of its being co-hyponymous with *Grenzwert*. The equivalent assigned to *Festlegung*, i.e. 'specification', must now be converted to verb and participial form before replacing *setting* in the structure found in Example 4, and *Grenzwerte* – of which, in the current database, there are only two instances, both rendered as 'limit values' – is inserted into the slot held by *Mindeststandard*, resulting in by specifying limit values.

These blocks (*fuel consumption, by specifying limit values, must be reduced*) are now re–assembled, in accordance with the pattern suggested by Example 3, resulting in 'fuel consumption must be reduced by specifying limit values'.

5. Conclusion

This paper has proposed a new type of dictionary for EBMT: one informed by linguistic, rather than extralinguistic principles, and has provided a brief outline of its application in a 'real–life' environment.

*This work has benefited from the programming support of Alan Kerrigan

Notes

[1] Contrary to the practices of other approaches within the example–based machine translation (EBMT) paradigm, the database consists of examples derived not from actual translations, but from *functionally equivalent* parallel texts, to which, however, a menu–driven interface allows modifications to be made on–line. Although, at the time of writing, this archive contains only some 4,000 examples, it is being constantly updated.

[2] See Sadler (1989:49), for a similar stance with regard to hierarchical structuring: "Hierarchy–based word matching often proved inadequate. Matches sometimes failed when intuitively they should have succeeded. For example, the verbs *to eat* and *to cook* are, intuitively, interrelated, but this is not reflected in the verb hierarchy (cook –> prepare –> make ready; eat –> ingest –> take in), for the simple reason that the relationship is not hierarchical but *procedural*."

[3] The coding for *schâdlich* illustrates that hyponymy/hyperonymy is taken to be one of the principal relations underlying the semantic organization of adjectives, which runs counter to Fellbaum et al. (1993:27), who argue that "it is not clear what it would mean to say that one adjective "is a kind of" some other adjective". The importance of this relationship in translation–oriented analogical reasoning procedures is, however, self–evident: if *Bestimmungen* in *forstrechtliche Bestimmungen* is rendered, in English, as "regulations", then, by analogy, *Bestimmungen* in *rechtliche Bestimmungen* can be assigned the same target–language (TL) correspondent.

[4] Synonymy, it should be noted, is viewed here as a symmetrical relation, not as a relation with one preferential direction.

[5] The dictionary currently contains approximately 30,000 word forms.

[6] So far, as the only "real–life" domain to which the thesaural approach has been applied is registering for conferences (Sumita et al. (1991; 1993)), this necessity has been avoided.

[7] Sedelow and Sedelow (1988:238) provide, however, a note of warning about such calculations of semantic similarity, which, in effect, involve thesaurus tree traversals. Such traversals, they point out, "are not consistent, nor are they always reliable indicators of word association, even though in many instances they can be used". Indeed, some calculations of similarity, as carried out in previous work, (e.g. as high as 0.5 between *read*and *buy*(Sato and Nagao (1990)), and as low as 0.2 between the Japanese equivalents for *vegetable*and *potato* (Sato (1991b)) are decidedly counter–intuitive.

[8] Tests as to the efficacy of this prioritization are currently being conducted.

[9] At present, a little over a quarter of the entries are phrasal in nature.

[10] See Thagard et al. (1990:264): "We contend that an analog is more likely to be retrieved the greater the degree of isomorphism it has with the structure that initiates the retrieval. Isomorphism is conceptually distinct from semantic similarity, in that two structures can be perfectly isomorphic even though they share no identical or similar elements."

11 Sometimes, complete semantic analogues are called up. 'Der Kauf bleifreien Benzins muß gefördert werden' was, for example, matched by 'der Verkauf bleifreien Benzins muß begünstigt werden', the dictionary allowing the synonymy between *fördern* and *begünstigen* and the antonymy between *Kauf* and *Verkauf* to be captured.

12 The example database suffers, as, for the time being at least, it is not informed by the results of domain modelling, from redundancy. See Sumita et al. (1993:89): "To generalize examples or to compress the example database with no drop in system performance is of importance from the standpoint of space and time complexity."

13 Alignment is intuitive (see Sato (1993:65)) and is performed interactively.

14 *Sentence–level* recombination has received relatively little attention in the EBMT literature. Sato, having given the lead (Sato and Nagao (1990)), has shifted his interest to lower–level units (1993), while Sumita et al. have been mainly concerned with nominal phrases (1991) and, more recently (1993), prepositional phrases.

15 Semantic correlates are used so that, if no match is found for *Kraftstoffverbrauch*, the structure of an analogue such as *Benzinverbrauch* ('petrol consumption') may be imitated, provided correspondents for *Kraftstoff* and *Verbrauch* are to be located, separately, in their simplex forms. This procedure has not been implemented yet, however.

16 Space constraints do not permit illustration of the way in which transformational relations are exploited at this point.

References

Agricola, E. 1972. *Semantische Relationen im Text und im System.* Mouton: The Hague 1972.

Fellbaum, C., Gross, D. and Miller, K. 1993. 'Adjectives in WordNet' in G.A. Miller et al., *Five Papers on WordNet.* Princeton University: Cognitive Science Laboratory Report 43.

Furuse, O. and Iida, H. 1992. 'Transfer–driven machine translation' in *Proceedings of International Workshop on Fundamental Research for the Future Generation of Natural Language Processing.* Manchester.

Michiels, A. and Noel, J. 1982. 'Approaches to thesaurus production' in *Proceedings of COLING–82,* Prague, 227–232.

Sadler, V. 1989. *Analogical Semantics: Disambiguation Techniques in DLT.* Dordrecht: Foris.

Sato, S. 1991a. 'Example–based translation approach' in *Proceedings of the International Workshop on Fundamental Research for the Future Generation of Natural Language Processing.* Kyoto: ATR Interpreting Research Laboratories, 1–16.

Sato, S. 1991b. *Example–Based Machine Translation.* Kyoto University, PhD thesis.

Sato, S. 1993. 'Example–based translation of technical terms' in *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation.* Kyoto, 58–68. Sato, S. and Nagao, M. 1990. 'Toward memory–based translation' in *Proceedings of COLING–90, Vol. 3, Helsinki,* 247–252.

Sedelow, S. and Sedelow, W. 1987. 'Semantic space'. *Computers and Translation,* 2: 235–245.

Sumita, E. and Iida, H. 1991. 'Experiments and prospects of example–based machine translation' in *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics.* 1991: 185–192.

Sumita, E., Furuse, O. and Iida, H. 1993. 'An example–based disambiguation of prepositional phrase attachment' in *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation.* Kyoto, 80–91.

Thagard, P., Holyoak, K., Nelson, G. and Gochfeld, D. 1990. 'Analog retrieval by constraint satisfaction'. *Artificial Intelligence,* 46: 259–310.