Types of Semantic Information Necessary in a Machine Translation Lexicon

David Mowatt

UMIST, Manchester, UK
dmowatt@fs1.ccl.umist.ac.uk | http://www.ccl.umist.ac.uk/ra/dmowatt/

Abstract

This paper describes research undertaken into assessing what types of semantic information (SI) are needed in a Machine Translation (MT) lexicon in order for 'good' translation quality to be attainable. We present a typology of semantic information, allowing the use of semantics in any MT system to be quantified in precise and absolute, rather than relative, terms. This typology was used to survey the SI present in twenty commercial and research MT systems. An automatically translated corpus was analysed to identify which *types* of semantics were necessary to achieve high quality translation. The survey and the analysis allowed us to conclude that four of the nine types of SI identified should *always* be included and that a further two complex SI types should be *considered* for inclusion pending further analysis. A formal lexicon specification incorporating these six SI types is presented.

1. Introduction

It is interesting to note that linguistic researchers have moved from primitive context free grammars (CFGs) to much more complex grammars such as HPSG, where the debate is not so much *what* **syntax** is encoded to *how* it is represented and how syntax interacts with other linguistic information, such as semantics and pragmatics.

The debate on *how much* or even which *types* of **semantic** information (henceforth SI) is needed (by rule-based natural language analysis systems) to disambiguate sentences has not however progressed to the same extent. The MultiLex project supports this assertion. The lexicon specification present within the report contained only the very *minimum* level of semantic coverage, "only the first step towards the building of a type system".[i] The authors agree that more semantics are needed, yet fail to provide an actual lexicon specification for them.[ii]

Our research aimed to provide a concrete lexicon specification for those wishing to use more complex semantics. In order to find out what types of SI the lexicon should be able to support, we first researched the existing semantic content of MT / NLP system lexicons as well as assessing the utility of each type.

2. Semantics in the Lexicon

The first aim of our research was to find out what *types* of SI should be included in MT systems – System ***Y*** uses SI types *3*, *4* and *8*, System ***Z*** just uses type *5*. It was *not* to propose an answer to the (perhaps impossible) question of just what quantity of SI should be included – System ***X*** should have ***n*** amount of SI type *1*, ***m*** amount of SI type *2*…[iii]. We examined

[i] Cf. McNaught and Smith (ed.) 1992:99.

[ii] The is also true for the extensive EAGLES survey of semantics, cf. Sanfillipo et al. (1998).

[iii] Nirenburg et al. (1996) did attempt this reduction of SI in a system to mathematical figures, though their methodology again only allows one to quantify SI in MT systems in a particularly imprecise manner.

semantics in existing MT systems and in lexicon documentation (such as MultiLex, Eurotra, Systran, KBMT) in order to draw up a typology of SI. While the author is certain that other SI types could exist, the typology presented does seem to successfully account for every full-scale (non-theoretical) system.

2.1. A Typology of Semantic Information

The table below contains a typology of SI. Each type is accompanied by a description and, in column (#), the number of commercial systems who claimed, when surveyed, to incorporate this type of SI. The order of the types was determined by their complexity (s**emantic primitives** are generally yes/no attributes, whereas **ontologies** are complex interrelated structures), how they are used (s**ubject domain** information provides *general indications* whereas **real world knowledge** (RWK) allow *complex semantic interpretations* of sentences) and how commonly a type was incorporated into systems featured in the survey:

t	*T*ype of Information	*E*xplanation	#
t_0	None	No semantic information is used at all	1
t_1	Sense Order	The senses are stored in order of frequency: Plant s_1 = non-animate living thing Plant s_2 = factory	1
t_2	Sense Frequencies	Proportional measurements of frequency Plant non-animate = 0.8, Plant factory = 0.2	10
t_3	Subject Domain	The typical subject area(s) of a given sense: Plant s_1 (living) = General Domain Plant s_2 (factory) = Manufacturing Domain	18
t_4	Semantic Primitives	A limited set of primitives: abstract/real, +/-place, +/-time: Plant s_1 = real, living, non-sentient Plant s_2 (factory) = real, place	14
t_5	Collocations	Words commonly found together. No grammatical relation between words identified. Plant (life), (manufacturing) plant	15
t_6	Common Subjects and Objects	As with collocations, but with a grammatical relation. “grow s_2 ***object*** = plant s_1”	12
t_7	Frames[iv] / Real World Knowledge	Uses not just the triple (WORD1 | GRAMMATICAL/ SEMANTIC RELATION | WORD2), but a complex interlinked relationship with lots of actions, properties, and different relations. The verb 'to plant' involves a **living_plant** object which is placed in the **ground** by a **human**, typically a **gardener** etc.	3
t_8	Use of an IS_A Ontology	Semantic primitives are used to form a complex Plant s_1 **IS_A** vegetative_lifeform Plant s_2 **IS_A** industrial_place	2 small / 1 big
t_9	Extended ontologies	Nodes in the ontology use more than the **IS_A** relation: Plant s_1 **LOCATION** garden Plant s_1 **HAS_PART** stem	0

[iv] Frames were used extensively in the research of Hirst (cf. Hirst, 1987) to semantically encode events, and all the subevents and concepts of which they are formed. Currently, only a few systems use descriptions like frames to encode semantic information, such as Fujitsu’s Atlas, and Transparent’s Transcend.

The typology presented is unique not because of the individual types identified, but because it the first attempt to present a method of identifying and classifying *all* types of SI with such a level of precision.

2.2. Semantic Information in Actual MT Systems

The survey of SI in the systems revealed both the vast majority of systems surveyed (70%) relied heavily on types $\boldsymbol{t}_4$ – $\boldsymbol{t}_6$ for disambiguation: semantic primitives ($\boldsymbol{t}_4$) and collocation information either with or without having identifying specific relations between words (types $\boldsymbol{t}_6$ and $\boldsymbol{t}_5$). The 'popularity' of these three types comes from their tremendous power to disambiguate senses (cf. for example, Yarowsky, 1995:189).

Globalink, GMS, LOGOS are amongst those systems which claim to use more sophisticated types of SI. Globalink's top of the range systems and GMS's T1 systems are of comparable quality – though notably, despite their commercial success, they are both said to be under-performing because the SI present in the systems is not complex enough[v]; LOGOS is also highly reputed, yet their system differs as customers are expected to tailor lexica and grammars much more than other systems demand.

However, counter to the belief that more complex SI will necessarily yield better results, Systran is frequently accepted to be the world's best performing MT system yet it only incorporates SI types $\boldsymbol{t}_1$ – $\boldsymbol{t}_6$, and some $\boldsymbol{t}_8$ The survey also revealed though that they will be incorporating more SI ($\boldsymbol{t}_2$) to further improve translation quality.

No-one knows how well a system incorporating all *nine* identified SI types would translate because no-one has ever successfully developed a full system which can use this quantity of SI. This would require, according to researchers of KBMT, "*all relevant domain knowledge*" (Nyberg, 1992:3) – a near impossibility therefore for free text input. Full $\boldsymbol{t}_1$-$\boldsymbol{t}_9$ systems certainly do not exist yet[vi]. Many have experimented with large levels of semantics, such as Memtah, only to abandon a full scale implementation in favour of a "only if needed" [vii] basis only. The KBMT team themselves admit that such systems are currently only financially feasible in very restricted domains, Nyberg (1997:2).

3. Using Translated Corpora to develop a Lexicon Specification

Based on the assumption that a translation of certain sentences is only possible with a certain level of SI, we wanted to discover which types of SI could be said to be necessary in an MT lexicon so that we could incorporate them into our lexicon specification. As we were particularly interested in translation for assimilation[viii], we also wanted to be able to say that we felt types $\boldsymbol{t}_x$ and $\boldsymbol{t}_y$ did not merit the effort required to input them into each lexical entry because they did not significantly enhance translation quality. We wanted to propose a lexicon specification that would provide enough information for disambiguation without being inherently financially infeasible.

A 3000 word bilingual corpus of texts on the subject of the World Wide Web, containing lots of domain specific terminology, technical language and general language, was translated by a human, by Systran Professional and by Globalink's French Assistant. Using knowledge about how MT systems worked, we attempted to deduce in each case *why* a system did not translate correctly (what we will term as a **reverse-perspective evaluation**). The error report thus contained:

[v] Ann Devitt, Lernout and Hauspie, during Trinity College Dublin Computational Linguistics Seminar Series.
[vi] Dahlgren, 1988, claimed to have a text-understanding system that used up to type $\boldsymbol{t}_9$, but with a vocabulary of only several thousand words, compared to the 360,000 words contained in Systran's lexicons
[vii] ipc. Helen McKay, Adacel Technologies, 1998
[viii] Cf. Jordan, Dorr and Benoit, 1993:50

(1) the area where changes need to be made (the lexicon, the grammar or both)
(2) whether or not the MT system would be able to successfully translate the phrase it got wrong if the new word / sense or new syntactic structure was added to the existing lexicon / grammar
(3) if the existing specification was insufficient, what would need to be added to the system for it to translate correctly ?

A preliminary examination of the corpora revealed that too many mistakes were made in the translations produced by the French Assistant software for there to have been value for the project in analysing them:

Systran: Les bruits et les films sont également possibles, cependant souvent trop grand pour que beaucoup de gens téléchargent et pour entendent ou regarder.

Globalink : Ses et cinéma sont aussi possibles, pourtant souvent trop grand télécharger pour beaucoup de personnes et entend ou vue.

All following observations thus relate only to the translation performed by Systran.

4. A Specification for the Semantics Section of an MT Lexicon

The results of the analysis were combined with knowledge of what SI types are already used in MT systems to propose a specification for the semantics section of an MT lexicon. Though many examples were used as evidence in the original analysis, we limit ourselves to just one example per SI type in this paper - each example has incorrectly translated words underlined, followed by the source language word and the correct translation in brackets.

4.1. Subject Domains and Sense Frequency - t_1, t_2 - and Subject Domains - t_3

We recommend that sense frequency and subject domain information should be included. In the following example identifying the IT domain would have ensured correct translation of three incorrectly translated words:

> From a technical point of view, the WWW connects <u>waiters</u> (*serveurs* = servers) HTTP who send pages HTML to <u>stations</u> (*postes* = terminal) equipped with a <u>navigator</u> (*navigateur* = browser).

Subject Domain = (Agriculture | Banking | Computers | etc.) *
Sense Frequency = `(0.0 – 1.0) ?`

A full list of values for **Subject Domain** could not be specified as only the IT domain was studied in this research, but we do though expect this field however to be dynamic in nature and expand as more domains are added. **Sense Frequency** is intended to be a 'last resort' mechanism when all other disambiguation fails: SI type t_1 is here made redundant by type t_2.

4.2. Inheritance, IS_A Ontologies and Semantic Primitives – t_4 and t_8

Virtually all MT systems use semantic primitives, though we found that using an IS_A ontology t_8 can replace them, be more powerful and still be both computationally efficient and easy to code into lexical entries.

> By Web you can visit an <u>exposure</u> (*exposition* = exhibition), read your newspaper, learn English, <u>control</u> (*commander* = order) a pizza pie.

Inherit = (Word#Sense) *

Stating all of the possible objects for *commander* or *visiter* would be time consuming and ad-hoc – using an IS_A ontology provides the same semantic power but with much greater ease. The top level ontology nodes can be used to encode features such as `sentience`, `abstractness` and `movability` and these can perform the same function as primitives (cf. Mowatt 1997 for a complete list). SI type $\boldsymbol{t}_4$ is thus made redundant by type $\boldsymbol{t}_8$.

4.3. *Real World Properties* - t_7 *and* t_9

RWK is generally accepted as being time-consuming to code and as *necessary* only when translating a *small* percentage of sentences, yet it is also accepted that using it sometimes provides the only way to translate certain sentences correctly.

En cliquant sur le nom de l'auteur d'un article vous pouvez ainsi avoir son adresse, sa photo…

While clicking on the name of the author of an article you can thus have his address, its (son == his) photograph…

"his photo" could either mean "the photo he possesses" or "the photo that pictures him". The `owner` of the photo would have to be a human or an institution (i.e. a gallery might own the photo), and the picture in the photo (the `content`) would have to be a tangible thing. If this was encoded under the RWK categories of `owner = sentient` and `content = real`, then only "author" would be a plausible candidate – "nom", "article" and "adresse" are ruled out.

RWProperty	=	**<RWFieldName> <RWFieldValue>**
RWFieldName	=	`IsA	HasPart	Content	Duration	(etc.)`
RWFieldValue	=	**"Word#Sense"+ | <Ontology_Node>+**
Ontology_Node	=	`Profession	Body_Part	Male	NoSize (etc.)`

Use of RWK cannot (yet) be definitively asserted as *necessary* for high quality translation, but we acknowledge its potential utility by including a possible way of encoding it in the lexicon whilst accepting that further research into RWK for large-scale systems is required.

4.4. *Selectional Preferences* – t_5 *&* t_6

Selectional preferences provide a way of stating the preferred subjects / objects etc. of a given word using semantic or collocation information. The correct translation of *mondial*, *universel*, *global* (in French or English) and *world* is most easily achieved by having a list of collocation objects, rather than complex semantic analysis:

Une dimension mondiale = A global (not world) dimension. Also: village **planétaire** = **global** village; guerre **mondiale** = **world** war, and similar words such as mortal / fatal / deadly.

SelPref	=	**<ApplyTo> <Field> <Operator> <Value>**
ApplyTo	=	`Subject	Object 1	Object 2	Modified Unit`
Field	=	**(Any lexicon field) | <RWField> | (Text | Word#Sense)**
Operator	=	`Equals	Not Equals`
Value	=	**"Text" | (Appropriate field value)**

The precise mechanics of *how*, given multiple selectional preferences for different senses, the parser is to weight each preference, is a problem that will not be discussed here. Our concern is to give the parser enough linguistic power and information to be theoretically able to disambiguate sentences, *not* to propose how it actually achieves it.

5. Conclusion and Further Work

We have presented a typology of semantic information which identifies nine individual types. When this typology was used in a survey of MT systems, it was revealed that no system currently in existence makes full use of sophisticated SI types such as real world knowledge $\boldsymbol{t}_7$ and extended ontologies $\boldsymbol{t}_9$, and few make use of IS_A ontologies $\boldsymbol{t}_8$.

Our analysis of the translated corpora suggested that of the nine SI types, $\boldsymbol{t}_2$, $\boldsymbol{t}_3$, and $\boldsymbol{t}_8$ should be incorporated into each lexical entry, and that type $\boldsymbol{t}_6$ should be used as often as is necessary. The analysis also strongly suggested that types t_7 and $\mathbf{t}_9$ would be of value to the translation process and for this reason we have proposed a possible implementation in the lexicon.

Our current research is focussing on what quantity of SI types t7 and t9 is needed (in each entry) to achieve translation quality that is unobtainable using the other seven types alone. The results of this research should be available towards the end of this year.

6. Acknowledgements

Thanks primarily to Blaise Nkwenti-Azeh who supervised the original MSc dissertation on which this paper was based and who later helped with the manuscript. Many thanks also to all of the researchers from the MT companies who participated in the survey of semantic types in MT systems. Finally, we would like to thank Systran for translating our test corpora.

7. References

COPELAND C., DURAND J., KRAUWER S., MAEGAARD B. Ed. (1991). *Studies in Machine Translation and Natural Language Processing : The Eurotra Linguistic Specifications.* Office for Official Publication of the Commission of the European Community. Luxembourg.

DAHLGREN K. (1988) *Naive Semantics for Natural Language Understanding.* Kluwer Publishers, Boston, Massachusetts.

GATES, D. ET AL. (1989) Lexicons. In *Machine Translation.* Vol. 4, No. 1, March 1989. pp.67-112

HIRST, G. (1987) Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press, Cambridge.

JORDAN P., DORR B. AND BENOIT J. (1993) A First-Pass Approach for Evaluating Machine Translation Systems *in the Journal of Machine Translation,* Vol. 6, Nos. 1-2. Kluwer Publishers, pp49-58

HUTCHINS W. AND SOMERS H. (1992) *An Introduction to Machine Translation.* Academic Press, London

MCNAUGHT J. AND SMITH S. Ed. (1992) *MultiLex : Definition of the Standard Monolingual Description of Lexical Items.* ESPRIT Project 5304 MultiLex

MOWATT D. (1997*) A Knowledge Rich Lexicon Specification for Transfer-Based Machine Translation Systems*. UMIST, Manchester.

NIRENBURG S., CARBONELL J., TOMITA M., GOODMAN K. (1992) *Machine Translation : A Knowledge-based Approach*. Morgan Kaufmann Publishers. San Mateo. California

NYBERG E. AND MITAMURA T. (1992) The KANT System: Fast, Accurate High Quality Translation in Practical Domains, in *Proceedings of COLING-92,* Nantes, France.

SADLER V. (1989) *Working with Analogical Semantics : Disambiguation Techniques in DLT.* Foris Publications, Dordrecht

SANFILIPPO A. ET AL (1998) *EAGLES Preliminary Recommendations on Semantic Encoding.* Interim Report *(not yet in public domain).*

YAROWSKY D. (1995) *Unsupervised Word Sense Disambiguation Methods Rivalling Supervised Methods.* In Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics [ACL95], Cambridge, MA. pp.189-196