GUÐRÚN MAGNÚSDÓTTIR

Collocations in Knowledge Based Machine Translation

Abstract

In order to generate colloquial language within the computational linguistics paradigm the problems of co-occurrence must be solved. As of yet research on co-occurrence has mostly focused on problems of syntax and selectional restrictions to describe the contextual relation within the sentence. Collocations and idioms have been neatly put aside as a unified problem to be dealt with in the lexicon or not at all.

In this paper collocations are defined according to the principles of semantics and a suggestion as to how to work on the retrieval of collocations, focussing on adjective noun constructions, from a text corpus will be made.

The research was carried out at the Center for Machine Translation, Carnegie Mellon University, together with Sergei Nirenburg and heavily inspired by Professor Allén's (Allén et al's 1975) work on collocations.

1 Defining Collocations

Idioms and collocations are two very different problems on a semantic level. The early definition of collocations (Firth 1957, Benson 1985) as being anything that frequently co-occurs can no longer be accepted. This definition discards the principle of syntactic atoms and would thus include such frequent patterns as 'it is' etc. Adding the constraint of atomicity would eliminate such patterns but would not be sufficient to distinguish between idioms and collocations.

Collocations are a string of words that co-occur under restrictions not definable by syntax nor selectional restrictions alone. These restrictions can be referred to as lexical restrictions since the selection of the lexical unit is not conceptual, thus synonyms cannot replace the collocate. The meaning of a collocation is compositional whereas the meaning of an idiom is not.

Collocations are compositional with hierarchical relations among the lexical units. The previous structural surface definition including a head, as the main word in the construction, and a collocate, as the supporting word is acceptable as is.

Proceedings of NODALIDA 1989, pages 204-207

2 Detecting Collocations in a Text

A multi-word idiom often violates selectional restrictions due to metaphorical use of words whereas a collocation will not. Thus an idiom may be detected in failing parses where a collocation will be parsed undetected. Collocations are 'permitted patterns' in contrast to idioms that are often 'prohibited patterns' within the selectional restriction frame.

Permitted pattern: 'large coke'

Prohibited pattern: 'as large as life'

The borderline between the two is difficult to draw. Such questions as is 'shining truth' a collocation or an idiom are indeed not easily answered. Objects that 'shine' have the property TO_REFLECT_LIGHT and are +CONCRETE. The property list of 'Truth' does not include these and should they be added it would result in extreme over-generation together with the possibility of faulty parses. Thus 'Truth' cannot 'shine' except in a idiomatic sense and will therefore be treated as an idiom.

Ambiguities may arise between an idiomatic meaning and non-idiomatic. Similarly, ambiguities may arise between a collocational meaning and non collocational meaning as in the example:

Decide on a boat.

Where 'on' is a preposition or a collocate. Out of context the sentence has two meanings and there is no way of deciding which is the right one. In such cases contextual information is the only disambiguating factor. Thus the manual labor involved when working on collocations will involve going through the corpus to detect the collocations as well as systematically entering them in a lexicon.

Retrieving collocations is not an easy task for a native speaker, simply due to the fact that a collocation is the natural way of expression that is more easily detected through violations in generation, in the output from a natural language system or a non-native speaker. It would be futile to provide a human user with the same interactive knowledge acquisition tool to work on both collocations and idioms.

Consider the sentences

There is a little light in the window.
There is a small light in the window.

The lemma 'light' has three different lexemes in Longman's, lexeme one, the noun, has sixteen senses. Of these the first five are most likely to appear in technical texts.

Light (cat NOUN)

sense 1: natural force U	property: QUANTITY
sense 2: source of light C	property: SIZE
sense 3: supply of light U	property: STRENGTH, QUALITY
sense 4: light (as time) U	property: QUANTITY
sense 5: set burning C	property: QUANTITY

The abbreviation 'U' stands for uncountable and 'C' for countable. The adjective little has four senses, linked to three scales.

Little—(cat ADJ)	
sense 1: small	scale: SIZE
sense 2: short - time	scale: TIME
sense 3: young	scale: AGE
sense 4: (idiomatic)	

The scale QUANTITY, which is probably the most frequently used meaning of 'little', is not present in the dictionary definition. It matches senses one, four, and five of 'light'.

The adjective 'small' has seven senses in Longman's. The last four have no relevance as to scaler meanings.

Small (cat ADJ)		
sense 1:	little in size,	scale: SIZE
	weight,	scale: WEIGHT
	force,	scale: STRENGTH
	importance	scale: IMPORTANCE
sense 2:	doing only a limited amount of X	scale: ACTIVITY
sense 3:	very little, slight (with U nouns)	scale: QUANTITY

Sense one of 'small' is very heavily compiled, for whatever reason. The scales of the two adjectives can be linked to the properties of the two concepts in a) and b).

a) Light ($IS-TOKEN-OF LIGHT)
b) Light ($IS-TOKEN-OF LIGHT-BULB)

Where a) represents sense one with the property QUANTITY, that given a modifier with a quantitative meaning, will give access to a certain position on the scale QUANTITY. In this case the position is represented by 'small' and 'little' as equivalent synonyms.

Sense two is represented by b) with the property SIZE that similarly gives the position on the scale SIZE, representing the equivalents 'small' and 'little'.

As for analysis this seems futile, since input texts are regarded as correct and the adjective is already present. However it is not possible to access an unambiguous result. Parsing at its best, and lexical mapping i.e. mapping surface expressions to concepts, will give two possible concepts, in a case where there is no ambiguity. Thus lexical restrictions would disambiguate between the concepts.

In generation the wrong choice of adjectives will lead to a wrong interpretation by the reader.

In order to be able to generate any information regarding these links the sentences need to be analyzed conceptually, that sort of analysis is only provided by knowledge based formalisms using ontologies and human interaction to ensure unambiguous results from the source language analysis.

3 Knowledge Based Machine Translation

At the Center for Machine Translation, Carnegie Mellon University, research has been carried out for some years on knowledge based machine translation. The most recent result is a prototype system (KBMT-89) delivered to the financier, IBM Japan, in february 1989.

The prototype system consists of different modules for natural language analysis, knowledge acquisition, and natural language generation. The system is an interlingua system relying on human interaction for disambiguation of multiple parsing results. The user is aided in the disambiguation process by the Augmentor, a specially built interaction component that allows the user to choose between the ambiguities at hand. The result is a meaning representation (interlingua) that is then the input of the generation component.

The analysis is based on a LFG like grammar together with the semantics present in the knowledge base or the ontology, the surface expressions are mapped on to the concepts in the ontology giving optimal grounds for knowledge acquisition.

Due to the fact that the result from the analysis and human interaction is unambiguous with links to the correct concepts, a filter for generation and disambiguation for the analyzed language can be generated. How this is to be systemized will be published in a forthcoming paper.

References

Allén, S. et al. 1975. *Nusvensk frekvensordbok baserad på tidningstext. 3. Ordförbindelser.* (Frequency Dictionary of Present-Day Swedish. 3. Collocations.) Stockholm.

Firth, J.R. 1957. *Papers in Linguistics 1934–1951.* Oxford: Oxford University Press.

Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press.

Cumming, S. 1987. *The Lexicon in Text Generation.* Dupl. 1987 Linguistic Institute Workshop: The Lexicon in Theoretical and Computational Perspective. Stanford University 1987.

Longman Dictionary of Contemporary English. 1987. Longman Group Ltd.

Magnúsdóttir, G. 1987. Fastcat Pilot-Study Report: Translation Systems and Translator Interviews. Språkdata.

Magnúsdóttir, G. 1988. Problems of Lexical Access in Machine Translation. In *Studies in Computer-Aided Lexicology.* Data Linguistica 18. Stockholm: Almqvist & Wiksell International.

Nirenburg et al. 1989. KBMT-89, project report. Center for Machine Translation. Carnegie Mellon University.

Department of Computational Linguistics,
University of Gothenburg,
412 98 Göteborg,
Sweden.
gudrun@hum.gu.se