Extraction of Concepts and Multilingual Information Schemes from French and English Economics Documents

Peggy Cadel, Hélène Ledouble

Laboratoire d'Ingénierie Linguistique et de Linguistique Appliquée Université de Nice-Sophia Antipolis {cadel;ledouble}@hermes.unice.fr

Abstract

This paper focuses on the linguistic analysis of economic information in French and English documents. Our objective is to establish domain-specific information schemes based on structural and conceptual information.

At the structural level, we define linguistic triggers that take into account each language's specificity. At the conceptual level, analysis of concepts and relations between concepts result in a classification, prior to the representation of schemes. The final outcome of this study is a mapping between linguistic and conceptual structures in the field of economics.

Introduction

Using two corpora issued from the Reuters database (600 articles in French and English), we give an outline on the creation of language-independent linguistic resources in economics. In the perspective of knowledge-based machine translation and information extraction applications¹, we show how the definition of a set of domain specific concepts combined to the description of structural information in both languages can help establish information schemes of recurrent economic information.

Concepts are selected and defined according to recurrent information using both corpora. Significant relations between concepts are specified according to their intrinsic and extrinsic properties. In parallel to this conceptual representation of information, linguistic structures of economic information are drawn from both corpora.

We first present the organisation of domain specific concepts into conceptual classes relevant to our application. In order to examine intralinguistic and interlinguistic knowledge associated to each type of information, we draw a parallel between the syntactic description of these structures in French and English and the semantic properties of the structures' components. We eventually give examples of mappings between structures.

Conceptual Classes

Our objective is to establish a set of domain specific concepts and classify them according to their behaviour in context. In order to keep most informative elements of our corpora, we chose to focus on nouns and verbs, as the most informative elements of language.

Using a morpho-syntactic dictionary for both languages, we extracted simple as well as complex lexical units from

both corpora². From this unordered list of terms, we defined series of domain-specific concepts having a similar function, with various designations in French and in English. The distribution of lexical units in conceptual classes was however specific to the each language: e.g. concepts expressed with verbs in English are nouns in French and vice versa. As we chose to take into consideration nominal as well as verbal structures, we thus established cross-category classes, i.e. a category-and language-independent classification of lexical units into conceptual classes.

Two first sets of recurrent concepts are constituted by the following conceptual classes:

- A set of economic agents, e.g. *company, bank, syndicates, firme, gouvernment, entreprise, etc.*
- A set of economic products, e.g. *price*, *dollar*, *taux*, *chiffre d'affaires*, *capital*, *rate*, etc.

These concepts are then grouped in more accurate classes, highly dependent on the application in view. We then distinguish two other conceptual classes in which the first set of concepts are interrelated:

- A set of economic operations, where an economic agent carries out an operation (operation potentially bearing on an economic product), e.g. *buy, export, merger, investissement, vendre, rachat, réduire,* etc.
- A set of economic "movements", where economic products undergo a variation, with no agent involved, e.g. *rise, drop, hike, increase, hausse, progresser, chuter*, etc.

We classified these concepts in a type hierarchy (*isa* relations), relevant to our domain of application:

Economic Agents



² ibid.

¹ For modelisation and implementation, see [Zinglé 1999]





The principle of representation adopted here is based on an interlingual structure (Nirenburg, 1992), manipulating these ontological concepts, considered as languageindependent. The hierarchical layout of concepts and associated inheritance mechanism enables the transition of properties between concepts.

Other movements

Starting from these four major categories of conceptual classes, we chose to emphasise the type of relations between concepts, the principle of which is illustrated in the following section.

Information Schemes

In our project, we focus on economic information and more precisely on how information is conveyed through both corpora. We first examine information about a *variation in turnover* and its various syntactic realisations. The following examples present the potential (but not indispensable) information associated to variation information through verbal or nominal structures.

Variation expressed through verbal structures:

<u>Eng</u>. But turnover will rise to 100 billion pesetas from 80.14 billion in 1997 <u>Fr</u>. Le chiffre d'affaires augmente de 8,2% à 2,32 milliards, contre 2,15 milliards.

Information associated to the variation in one or both examples (in bold):

- The type of the variation: *upward movement*
- Quantity of variation: *expressed in percent*
- Final value of product after variation: *the value is followed by a monetary unit*
- Value of product before variation: *value followed by a monetary unit*

³ it could be opposed that some strategical operations can also be financial and vice versa, but the constitution of those classes is in close relation to our objectives. Variation expressed through nominal structures:

<u>Eng.</u> The profits rise came on a **19 percent** increase in turnover to **1**,552 million shillings. <u>Fr</u>. Sanofi, qui a fortement progressé récemment, perd ainsi 1,98% à 843 FF en raison de la déception causée par le **repli du chiffre d'affaires** du groupe au troisième trimestre 1998.

Information associated to the variation in one or both examples (in bold).

- The type of the variation: *upward / downward movement*
- Quantity of variation: expressed in percent
- Final value of product after variation: *value followed by a monetary unit*

We then study structural and conceptual information related to different types of information.

Structural information

Lexical and grammatical elements help us identify informative elements associated to the variation in turnover.

The quantity of variation (in percent) was introduced in our examples by the preposition *de* in French, and *on a* in English. In French, information about the evaluation (quantity) of the variation implies the use of the preposition *de*, be it preceded by a noun or a verb. We can also have a lexical gap (e.g. *le chiffre d'affaires gagne 15%*). In English, several prepositions convey the quantity of variation (*of, by,* etc.), as well as a lexical gap, but the preposition changes with verbs or nouns. The preposition *of* for instance, will be the used with nouns only (e.g. *a turnover increase of 13%*), whereas *by* will be more frequently found in verbal structures.

Besides, a verb or a noun indicating variation that is followed by \dot{a} and *contre* in French will respectively give information about the final value of the product and the initial value. In English, the same information is given by the variation element followed by *to* and *from*.

For each type of movement expressed by a specific lexical element, be it a verb or a noun, the following information about the structure can be gathered:

	Quantity	Initial value	Final value
Lexical	E. = percent	E. = shilling	E. = shilling
Element ⁴	F. = %	F. = dollar	F.= dollar
Grammatical	E.=on a, by, of	E. = from	$E_{\cdot} = to$
Element	F. = de	F. = contre	F. = à

⁴ The list of lexical elements is not exhaustive. Besides, lexical elements given for initial and final values are not systematically expressed (c.f. example in French)

Conceptual information

The relations between different types of concepts serve as constraints in the constitution of information.

This following conceptual graph⁵ based on concept types represents the maximal default information structure (with the whole of potential units) about turnover variation although all units of information are not systematically expressed in corpus.

 $[movement] - (isa) - [type] - (expr) - [economic_product]⁶ - (rslt) - [value] - (val) - [quantity]$

This representation, considered as language independent, takes into account the polymorphism of information (cf. examples above).

With a similar analysis of linguistic structures, we drew conceptual representations of various types of information, to which correspond a number of syntactic realisations in both languages:

Eng.: Krung Thai Bank, Thai Farmers Bank and Siam Commercial Bankall announced that they would slash rates by up to 150 basis points this month.

<u>Fr</u>. Les marchés européens dans leur ensemble ont bénéficié des décisions de **banques centrales espagnole, portugaise et suédoise** de **réduire leurs taux**, certains attendant que la Banque d'Angleterre en fasse de même prochainement.

A comprehensive list of linguistic structures matching a conceptual representation can be enumerated for each scheme. These examples show that some relations are essential to a given conceptual representation (e.g. the object (obj) relation above), whereas others (e.g. the value (val) relation) remain optional. For the concept TRADE, the default information scheme shows various potential relations as well:

[trade]	– (agnt)– [actor]		
	-(obj) - [product]		
	- (val) -[value]		
	- (dest) -[actor]		

Eng. [...] but only last month 11 company insiders were selling shares of the busines

Eng. All of its proprietary tobacco product lines, including the three premium **brands it is selling to Philip** Morris for \$300 million, include an additional warning statement

<u>Fr</u>. L'essentiel de la trésorerie disponible sera affecté à de nouveaux investissements de diversification et au programme de **rachat d'actions.**

<u>Fr</u>. Moulinex envisage de racheter 1% de son capital

The canonical graph defining the TRADE concept specifies the generic concept of *product* for the object relation, the instantiation of which can be different types of products (*economic* and *non economic*) defined in the hierarchy (*capital, company, shares*, etc.)

Besides, the above example shows an interesting treatment of the elements *brand* and *Moulinex*: these companies, that constitute the object of an exchange in those examples, are no longer considered as *agents*, but are classified in the conceptual class of *economic products* as well. Conceptual representations have to take into account those various perspectives in the description of information.

Perspectives

The conceptual representation of information that takes into account syntactic and semantic constraints is a recquisite "understanding" of documents prior to an automatic processing. The meaning of information can therefore be drawn not merely from the succession of linguistic elements in a sentence, but also from discontinued structures of information. Besides, these resources facilitate the localisation of recurring information under its different forms.

This study lies within the framework of knowledge-based translation or knowledge-based information extraction with a view to automatising some of these tasks. A parallel mapping with other languages would help build a more comprehensive multilingual knowledge base of domain-specific information. The relationship between interlinguistic and intralinguistic knowledge being particularly emphasised, both the reusability of knowledge and its availability within mono- and multilingual contexts is envisaged.

References

Cadel Peggy, Ledouble Hélène, 1999, « Elaboration d'une ontologie pour le traitement de documents automatiques bilingues », In JILA'99 (p291), Nice.

Cadel Peggy, 1999, Recherche d'informations et analyse syntactico-sémantique : un exemple avec le verbe progresser dans des documents économiques. In JILA'99 (pp65-67), Nice.

Fargues J. (1989). Des graphes pour coder le sens des phrases. Pour la science n°137. 52 - 60

Fargues J. (1989). Un système d'analyse de textes et de question / réponse basé sur les Graphes Conceptuels, Traitement des langues naturelles (p.69-76), 2ème Ecole tenue à l'ENSSAT-Lanion, 3-7 juillet 1989.

⁵ for the formalism and definition of relations, cf. [Sowa:1984]

 $^{^{\}rm 6}$ we consider the experiencer of this movement to be the economic product, even if it is its value that experiences a movement

Gruber, Thomas R., (1989). The acquisition of Strategic knowledge, Perspectives in artificial intelligence, vol. 4, Academic Press, INC. Th. Ph. Diss: University of Massachusetts

Guarino (N.), Giaretta (P.), (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. In Mars (N.J.I), Towards Very Large knowledge Bases: Knowledge Building and Knowledge Sharing (p.25-32). IOS Press.

Ledouble, Hélène (1998). Méthode d'analyse de documents économiques anglais en vue d'un traitement automatique. In Travaux du LILLA. n°3(pp 79-91), Nice.

Nirenburg S. et al. (1992) Machine Translation: A Knowlege-based Approach. Morgan Kaufman Publishers, California.

Sowa, John (1984). Conceptual Structures : Information Processing in Man and Machine. Addison Wesley Publishing Company.

Zinglé, Henri (1999). La modélisation des langues naturelles. Aspects théoriques et pratiques. Travaux du LILLA numéro spécial, Nice.