A TEXT-MEANING-TEXT DICTIONARY AND PROCESS

Dominique Dutoit

MEMODATA 17 rue Dumont d'Urville 14000 Caen, FRANCE memodata@wanadoo.fr Université de Caen Esplanade la Paix 14000 Caen, FRANCE dutoit@info.unicaen.fr

Abstract

In this article we deal with various applications of a multilingual semantic network named *The Integral Dictionary*. We revise different commercial applications that uses semantic networks and we show the results with the Integral Dictionary. The details of the semantic calculations are not given here but we show that contrary to the WordNet semantic net, the Integral Dictionary provides most data and relations needed to these calculations. The article presents results and discussion on lexical expanding, lexical reduction, WSD, query expansion, lexical translation, information extraction, document summary Emails sorting, catalogue access and information retrieval. We conclude that a resource like the Integral Dictionary can become a good new step for all those who tried to compute semantics with WordNet and that complementary between the two dictionaries could be seriously study in a shared project.

Introduction

In this paper we discuss the different applications of *The Integral Dictionary* (TID). TID is a multilingual semantic network. Based on a original database design that merges three semantic theories, we revise some real applications and we foresee others.

TID modules are thought to represent lexical semantic information and to carry out different functions that support semantic processes to perform most tasks in Natural Language Processing and applications in knowledge management.

TID database is basically structured as a merging of three semantic models available for five languages. The maximal coverage is given for the French language with 185.000 word-meanings in the database. English Language appears like the second language in term of coverage with 79.000 word-meanings. Three additional languages (Spanish, Italian and German) are present for about 39.500 senses. These smallest dictionaries with universal identifiers to assure the translation define the Basic Multilingual Dictionary available from the ELRA. A corpus coverage evaluation for the Basic Multilingual Dictionary has been done by Grefenstette (1998) with the newspapers corpora that has been used in the US-Text Retrieval Conference government-sponsored (TREC). The result was that the chance of pulling a random noun out of the different corpus was on average 92%. This statistic is given for the Basic Multilingual Dictionary and, of course, the French Integral Dictionary reaches the highest coverage. The three semantic models that organize TID are described below:

- A componential architecture where the ontology as WordNet or EuroWordNet is a particular artifact. In this componential architecture, we find three main ontologies for different purposes and a thematic net based on the lexicon justified collocations. The three main ontologies are given to organize
 - each specific linguistic ontology (for example, "relative to" is a big root for French but not for English). These specific linguistic ontologies are given for each Lexical Function (LF) in the sense of sense—text theory,
 - verbal ontology to organize the lexicon by grammatical categories like WordNet,

 semantic ontology to organize the lexicon in consideration with the world knowledge.

The thematic net based on the lexical justified collocations contains mainly:

- a componential semantics that organizes lexical sense by the presence of a needed or optional seme. The componential semantics incorporates the linguistic concept of differential semantics.
- an encyclopaedic organization used to describe concrete things or activities.
- 2. A world knowledge architecture of concepts where we can describe some common sense knowledge (it 's the area of CYC point of view (Lenat 1999)). In this area, each concept is syntactically and semantically defined. For example, we find here the whole of what happens for each word-meaning that has like hyperonym "animated", "events", "act". Note that we can show that these informations are not redundant with the thematic net because here the syntactical constraints like Entity-Predicate are given.
- 3. A derivational set of relations that provides the usual lexical functions (LF) used by each language. This set of relations is compatible with the sense—text theory (Mel'cuk 1992). The main use of these relations is to compute paraphrases from a given deep sense or to compare two sentences in consideration with the deep syntactical dependencies.

The Integral dictionary is different from traditional lexical resources and also different from similar types of resources because TID incorporates a lot of theoretical points of lexical description and of description of concepts. We show how this resource is running by the Semiograph to improve several NLP services and applications. In conclusion, we argue about the interest of a single semantic network that manages all the semantic informations needed in various linguistic processing.

The linguistic services

The linguistic services highly benefits from The Integral Dictionary because the richness of its descriptions. We group as follows the linguistic services:

- The linguistic services that facilitates the natural language understanding by human
- The linguistic services that facilitates the human writing

Human understanding help

These services are useful to get senses or lexical translations of an unknown word. The mechanism of this help may incorporate or not the context of the unknown word. If we consider the context, a word-sense disambiguation (WSD) task must precede the production of results. WSD will be studied later in this article, so we will not use here the contextual information. In this case, we study only the help services for human understanding. We know that this service is well done if the system provide definitions, domains of use, synonyms and other types of information. The Integral Dictionary organizes the lexical sense by a set of 40.000 concepts or seme for the complete dictionary. Contrary to WordNet that knows as many synsets as definitions of synset, the Integral Dictionary provide word-sense by a formal generation of 1 to i structured concepts. For a sample example, TDI don't say together that

a florist is a person that sells cut flowers or plants for inside the house

but that

a florist is a (Is-a) \seller\\1 or \shop assistant\

is concerned by \text{\text{trading}\ (like the seller)}
is concerned by \text{\text{flower}\ and \text{\text{plant}\ }
is the tautology of \text{person} (subject)
+ \sell\ (verb)

+\flower and plant\ (object)

Because each TDI concepts can show their whole contents, the user that uses TDI may see automatically florist in the context of selected analogies. For example the list of the 234 sellers (statistics is given for French), the thematic net of \tangle trading \tau, the complete list of flowers or plants inside the houses... It's equally possible to show the florist in the complete context of a person if the word seller is unknown by the user. Finally, it's possible to get a translation of the term florist by two way:

- By following a link translated to we obtain the French word fleuriste
- Sometimes this link is not available because polysemy depends to a particular language. In this case, an operation of TDI computes the best translations by comparing together the two sets of semantic concepts (see below semantic distance).

Note that The Integral Dictionary contains 700.000 links between word-senses and concepts (for French) and around 90.000 links between concepts. It exists 129 kinds of relations (such as *Is-a*) and 48 kinds of concepts (such as \seller (*list*)\).

Human writing help

By human writing help we can think about ideas to words dictionary. The Roget is a good example of this type of dictionary. However, the organization of Roget is sometimes too coarse to suggest without any additional verification the good term for an idea. A complete specific sense—text (Mel'cuk 1992) functions may be used. The sense—text theory suggests the use of about a hundred of lexical functions (LF). The Integral Dictionary may provide some word-sense expansion using these relations. For example, for the verb to sell we find:

Action(to sell)

→sale

Result(to sell) →sale
Converse123(to sell) →to buy

Agent(to sell) →\ seller, salesman \
Able(to sell) → saleable, marketable
Unable(to sell) → unsaleable, unmarketable
Object(to sell) →\ products and services \
Beneficiary(to sell)= Agent(Converse123(to sell))

→buyer, purchaser

Etc.

The calculated linguistic services

By calculated linguistic services we consider the linguistic services that a sample reading of the dictionary can not provide. It happens when it's impossible to define a concrete lexical function. In sense-text theory, it's possible to find a LF as to cry out #for an animal# or as a money #for a State# and so on. Sense-text theory considers a large number of this particular lexical functions. However we consider that these lexical functions are too particular and too limited with a defined number of arguments. So we consider that these lexical functions may be calculated by the graph itself. The general idea of the algorithm is that each generic term (each term that knows specific terms) is the possible release of this calculation of specific terms. So, when we consider a generic term (or hyperonym) we consider together all of its specific terms that can be activated by the context. At the end of the calculation, we can substitute the generic term and the activating context by the specific term. We give some examples of this. The first examples come from a particular API of the Integral Dictionary that resolves the search of words by definitions of them. It's a particular application of natural language generation (NLG). The second examples come from concrete texts. The general term for this NLG operation is (lexical) reduction.

Reduction with the inverted dictionary

We write in italics the generic terms.

Query		Results
To make the sound of a male turkey Crier pour un dindon	\rightarrow	To goggle Glouglouter
Monetary unit of Japan Monnaie du Japon	\rightarrow	Yen Yen
Seller of flowers Vendeur de fleurs	\rightarrow	Fleurist Fleuriste
Seller of magnolia Vendeur de magnolia	\rightarrow	Fleurist Fleuriste
Person that sells cut flowers or plants for inside the house Personne qui vend des fleurs coupées ou des plantes d'intérieur	→	Fleurist Fleuriste
Person that sells tulip Personne qui vend des tulipes	\rightarrow	Fleurist Fleuriste
Person who sails on the sea and uses force to steal from other ships Personne qui parcourt les mers en vue de piller les navires.	→	Pirate Pirate
Person whose job is treating people's teeth Médecin des dents	\rightarrow	Dentist Dentiste
To say that something is not true Délcarer faux qqch	\rightarrow	To deny Dénier

¹ The concepts are written with the \...\ marks of parentheses.

Query		Results
Pain in the lower part of the back Douleurs dans le bas du dos	\rightarrow	Lumbago Lumbago

Table 2: Reduction with the inverted dictionary

Reduction from concrete texts

The examples given below was founded in literature or economic newspaper.

The money of Tokyo La monnaie du Pays du Soleil Levant La monnaie d'Hiro-Hito	→	Yen Yen
Les forestiers coupaient	\rightarrow	Abattre
Un sentiment de colère	\rightarrow	Colère
Le successeur de Saint Pierre	\rightarrow	Le Pape
La voiture du Pape	\rightarrow	Papamobile
La capitale française	\rightarrow	Paris
La <i>capitale</i> française (with a popular denomination)		Paname
Les habitants d'Ile de France	\rightarrow	Francilien

Table 2: Reduction from concrete texts

Contrary to the given results, generally we can find more than only one answer for each reduction. In fact, multiple answers are given in accordance with an analogical order. For example, a phrase like *la capitale française* gives a complete classification of the capitals in accordance with the distance between the France and the other countries. For the example *Person who sails on the sea and uses force to steal from other ships*, we get different analogically classified pirates as the figure below shows:

	per se entire harman was seen in	estato)		State of the last	Calcu	•
Max octivé 50	10	tanca age limits	irisya s	CEIAIN HON	oforguna 5 Fichi chieration 1	br
clé li	mité libellé	OTHER WAY	activa	tion	origine scoreD mots a	ctiv
nor n°1 e	genturier: guntave :	PHDASE I	METNALE .	·· TAVOR	position : 0 à 10; négat	ion
13137	aventurier	1786	F	0.0	(-1,-1,-1,-1,1422)	
117911	frère de la côte	1267	S	6.129	(-1,-1,1197,1286,1320)	
13140	pirate	1267	s	6.129	(-1,-1,1197,1286,1320)	
90438	boucanier	1267	s	6.129	(-1,-1,1197,1286,1320)	
118408	corsaire barbaresq		s	7.276	(-1,-1,1174,-1,1298)	
13139	flibustier	1267	5	7.412	(-1,-1,1197,1286,1320)	
23312	forban	1255	S	7.560	(-1,-1,1197,1251,1320)	į.
80006	océensute	1392	S	8.503	(-1,-1,1100,-1,1224)	ă.
80007	aguenauce	1392	S	8.503	(-1,-1,1100,-1,1224)	
160764	reis	1450	s	8.545	(-1,-1,1174,-1,1298)	
117828	corpaire	1450	5	8.757	(-1,-1,1174,-1,1298)	
not n°2 p	arcourt; syntaxe : PE	PASE_NO	THALF; RE	LATIVE_O	UI; VERBE; VERBE; MOYAU; VER	BE;
38533	parcourir	2000	P	0.0	(-1,-1,-1,-1,-1)	
18772	parcourir	2000	P	0.0	(-1,-1,-1,-1,-1)	
			LE ; PELATI		EPBE; VERBE; NOYAU; COD; NOY	AU:
27250	ner	2000	P	0.0	(-1,-1,-1,-1,-1)	
€ (2402/243)	美国新疆民众政治发展的发展的	2 15 2 2			SECTION OF THE PROPERTY OF THE	等10

Figure 1: An example of reduction

In English, for the same query, the results are:

Meaning identifier	Word-sense	Score by activation	Score by complete analysis
131140	Pirate	1267	6.129
90438	Buccaneer	1267	6.129
131139	Freebooter	1267	7.412
117828	Corsair	1450	8.757

Table 3: Results in English

Lower the scores are, better are the results. For the activation measure, the scores are contained between 1200, the minimum, and 2000 the maximum (no activation). For the complete analysis score, the scores are contained between 0 (identity), and 140 (complete difference). Activation and complete analysis give not exactly the same results because in one case the best links are used and in the second case all the semantic links are evaluated. So, complete analysis uses some semantic constraints that the partial analysis doesn't see. The two algorithms of semantic distance (not detailed here) are used differently in accordance with the linguistic application needs.

The linguistic application services

By linguistic application services we think about information retrieval, information filtering, lexical expansion, lexical translation, document summary, Emails sorting and answer, catalogue access and so on. We see that all these semantic applications need a common basic algorithm: it's the word sense disambiguation process (WSD).

The Word Sense Disambiguation

To select the best reduction of a query to the inverted dictionary, the previous application used two WSD algorithms. The first one was based on the activation. The second one was based on complete semantic closeness. In TID the distances between words or between word an a complete sentence or between two sentences are a function of the graph of concepts. TID is an acyclic graph (and not a tree) with a common root named *Dicologic*.

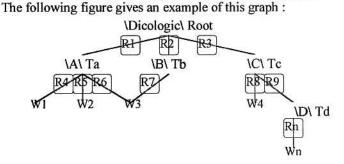


Figure 2: The Integral Dictionary

In this figure, we read that $\Dicologic\$, a root, has a descendant labelled $\A\$ that is a Ta concept with a kind of link RI. Vice versa, we read that the concept $\A\$, a Ta concept, has an antecedent labelled $\Dicologic\$ that is a root by a RI^{-1} kind of relation. We suppose that the knots labeled \A i are words (not necessary final words...).

We see that W3 has two antecedents by two kinds of links. These antecedents are: 1.41 and 1.81. Mostly, words and concepts of the Integral Dictionary have more than only one antecedent. TID knows 40.000 concepts and around 185.000 word-senses for French. The average depth of the graph is 15. The average number of direct ascendant is 2.1. The average number of belonging to a concept for a word is $15^{2.1}$ =294. The average number of word-senses that are reachable from a given word-sense is 6.500 (for French).

The activation distance between W1 and W3 is a function of |A| and the kind or relations R4 and R6. The complete distance between W1 and W3 is a function of |A| (with R4 and R6) and Dicologic (with R4+R1, R7+R2).

By this way, the activation distance between W1 and W3 is the same of the activation distance between W1 and W2 if the kinds of relation are similar. At the opposite, the complete distance between W1 and W2 is smaller than the complete distance between W1 and W3. This result is a direct application of differential semantics.

These previous propositions are true if the gotten concepts are syntactically neutral. If not, the results may be different.

With the complete graph, to get the criteria of semantic measure is quite complex. We will describe elsewhere optimal algorithm to manage massive poly-hierarchical graph and to compute the semantic measure. In this paper, we provide only final results : with the fine granularity of TID, the WSD algorithm, that uses only the TID data, reaches a percentage of 65% of good and unique disambiguation (without any other technique or data). The percent of success was nearly the same from 1992 (Dutoit 1992). This level of result is near the maximum possible because around 20 % of the lexicon is polysemous and supposes multiple answers. Note that disambiguations are uncommon: 1 to 5 % in dependence with the corpora.

The application services

All of these following applications are java API's of our development tool named *the Semiograph*². In this paragraph, most of the examples will be give in French Language.

Query expansion

This application gives an help to the users who put a query to a search engine. In this area, the Semiograph has to determinate the sense of the query and generate (or suggest a generation) an expansion of the query in accordance to the syntactical properties of the source.

The Semiograph links independent mechanisms of expansion defined by the user. Eight mechanisms are available:

- Alias: to get the graphics variant
- · Synonyms: to get synonyms for a meaning
- Hyperonyms (level): to get hyperonyms for a meaning
- Hyponyms (level): to get hyponyms for a meaning
- Inflected forms: to get the inflected for a meaning
- Derived forms (boolean: syntax): to get correct lexical functions in accordance or not with the syntactical proposition
- Geographical belonging (level, level-1): to get toponyms
- Translation (language parameter): to get a translation of the query.

Figure 2: An *Alias;derived*|*F;synonyms* expansion Of course, the useful parameter of this application is, for the French language, the parameter *Inflected form*. We draw attention to the boolean *sens unique seulement*. This boolean limits the expansion to the unambiguous words.

Lexical translation

In French, *brouillard* is an homonymous word. The translation of the common sense is *fog*. An uncommon sense is *daybook*. The sentence:

Georges a imprimé le bilan avant de vérifier le brouillard may be lexically translated by :

Georges printed the statement of accounts before checking the daybook.

At the moment, the most popular machine translation suggests wrong translations.

Systran³ generates:

George A printed paper form the assessment before of check the fog.

Lernout an Hauspie⁴ suggests (after selecting the topic vocabulary of *business*):

Georges printed the balance sheet before verifying fog. The help of dictionary like TID may improve machine translation in by selecting correct lexical translation.

Information extraction

The MUC conference shows multiple applications of WordNet to define template for information extraction. Obviously, the use of TID is the same to WordNet in this area.

Document summary

Different approaches exists in the area of Document summary. The most popular way seems to be based on syntactic features. Lexical semantics provides a new approach. Generally, the topics of a document are contained by word-senses particularly well disambiguated by the context. If a document treats of billiards, the context of billiards is necessarily saturated by terms of larger topics like \games\,\competition\,\dexterity\\(\(\)\((\)\)\ and terms in dependence with \billiard\\\ billiard\\\ like \ball, \cue, \cannon\). Using this property, lexical semantics defines

Expansion d'une phrase à partir d'un script aboulissement d'un projet de voiture électrique ALIAS, DERIVED | F, SYNONYMS iques seulement pour l'indexistion 32183 : 32183 : aboutissement : 0 : P : Jaucunel 185348 ; 32183 ; aboutir ; 10728 ; D ; [fait et résultat de] mot nº 2-70230; 70230; projet; 0; P; [aucune] 10910 ; 10910 ; projet ; 0 ; P ; [aucune] 98405 : 98405 : projet : 0 : P : laucunel 117616; 117616; projet; 0; P; [aucune] mot nº 3 83530 ; 83530 ; volture ; 0 ; P ; [aucune] 190865 ; 190865 ; volture ; 0 ; P ; [aucune] 184762 ; 184762 ; volture automobile ; 0 ; P ; [aucune] 32030 ; 184762 ; volture ; -19000 ; Y ; [Syn. de (-1)] 40209 ; 184762 ; automobile ; -19000 ; Y ; [Syn. de (-1)] 58245 ; 184762 ; auto ; -19000 ; Y ; [Syn. de (-1)] nontres 16 Mements inches and me

² POS tagging, syntactical analysis for French and WSD are also API of the Semiograph.

³ Accessible by the link translate of Alta-Vista

⁴ To try, use http://officeupdate.lhsl.com

lexical topics as the terms of a text that have a minimal complete distance between them and the rest of the text. Notes that this property may be used to verify the relevance of keywords manually given by a writer. An application may be the struggle to the spamming of search engine.

To give an example of result of lexical summary, the algorithm applied to this paper provides in the 20 best words the terms: lexicon, dictionary, semantic network, semantics, calculations, and disambiguation.

Emails sorting and answer

In this application, we have to sort a flow of documents according to a classification. Most systems execute this task after a learning phase. Learning phase poses problems because it needs a costly preliminary manual tagging of the documents. So, it's attractive to verify if a complex lexicon may provide the sorting without any learning.

To give an example of the result, we begin by determinate mailboxes. The formalism is very light: firstly, we define an identifier for each mailbox, secondly we define a definition of this mailbox. On the following examples, identifiers are given between parentheses and definitions are given after.

[guerre du Kosovo] guerre du Kosovo [tabac et jeunesse] tabac et jeunesse [alcoolisme et Bretagne] alcoolisme et Bretagne [investissement immobilier en Île-de-France] achat, vente et marché immobilier en Île-de-France

The definitions may be given in English with the exactly same result. The following text:

Les loyers stagnent à Paris mais la baisse de la TVA sur les dépenses de réparation de l'habitat devrait soutenir le marché de l'ancien

gives in term of complete distance:

[guerre du Kosovo]	135
[tabac et jeunesse]	140
[alcoolisme et Bretagne]	129
[investissement immobilier en Ile-de-France]	9

We observe that differences between the mailboxes are very marked (we remind that the best score is the lowest). Notes that this approach may be used to help the classifying of web sites that is today entirely manually carry out.

Catalogue access

This application is nearly the same one of the previous one. The main difference is that the document to compare are very small is this application: document size is given by the longer of a query.

Information retrieval

WSD and lexical summary may help us to improve the strategy of text indexing. Current strategy is given by complete full-text technique. A progress may be done by indexing only relevant terms with the criteria of lexical summary. The index size would be scaled down and the index would be more relevant of the content. WSD may improve too the quality of indexing if we suggest a way to memorize sense in context. The way that we suggest is to

add for each indexed terms some other terms positively chosen to allow query expansion with semantic constraint. For example, suppose that we add to each indexed terms, one of its generic, one synonym and one translation. In this case, a document indexed with the word brouillard (fog or daybook) will be no longer indexed by brouillard but by, for example, brouillard+(hyper=phénomène climatique, syn=purée de pois, Engl=fog...). This technique may reduce the noise and must be evaluated. But the main goal of semantics in the domain of information retrieval is not given by these nearly mature techniques.

We define that the goal of semantics in information retrieval is to maintain an association between the textual form in which knowledge is expressed in source documents, and an expressive narrative knowledge representation language. This goal may refer to deep semantic representation which is, beyond sample WSD, the challenge of the next future research.

Conclusion

The Integral Dictionary provides self-sufficient data for the most of common semantic calculations. Historically, TID and WordNet was built in parallel with different objectives. TID was precisely built to facilitate WSD since 1987. WordNet was built to show a new formal classification of the lexicon. We can show that these two points of view are complementary rather than concurrent. As a French partner for EuroWordNet, we suggest to merge the best ideas of these two semantic networks. We know that some research worker are interested by this proposal. Problem is given by the poor public research credit in the strategic and costly area of semantic resource building. Institutions wish applications. Industrial builders need more and better semantic resources. Because we are industrial, we need more and better resources. And we continue strongly the work in this area. In our domain, the quality of the semantic network is the main key. The key for best stand-alone semantic calculations. The key to improve the performance of data mining. The key to prepare the conjunction of data mining and stand-alone semantics. To conclude this conclusion, firstly, we claim that TID and Semiograph are available for research purpose for a cost depending to the cost of the transfert of technological, secondly, that we are attentive to develop international partnership.

References

- D. Dutoit 1992, A set theoritic approach to lexical semantics, Coling.
- C. Fellbaum 1998, WordNet: An Electronic Lexical Database, edited by Fellbaum, M.I.T. press.
- G. Grefenstette 1998, Evaluating the adequancy of a multilingual transfer dictionary for the Cross Language Information Retrieval, LREC 1998.
- D. B. Lenat 1999, From 2001 to 2001: Common Sense and the Mind of HAL, www.cyc.com/halslegacy.html.
- Mel'cuk 1992, Dictionnaire Explicatif et Combinatoire (DEC), III (Les presses de l'Université de Montréal).
- S. D. Richardson, W. B. Dolan, L. Vanderwende, 1998, Mindnet: acquiring and structuring semantic information from text, Coling.
- Y. Wilks, M. Stevenson 1998, WSD using Optimised Combinations of knowledges sources, Coling.