Natural Language Asymmetry and Internet Infrastructures*

Anna Maria Di Sciullo
Université du Québec à Montréal

Abstract

We present the main features of an Information Retrieval and Extraction system based on natural language asymmetric relations. We show that, along with the identification of functional elements, asymmetric relations contribute to improve the performance of search engines. We compare an Information Retrieval and Extraction system based on the recovery of a subset of asymmetric relations with current operating search engines based on key word search and Boolean analysis. We show the superiority of the first system. We show that natural language asymmetries constitute a crucial ingredient of Internet Infrastructures ensuring greater precision to internet communication.

1. Internet infrastructures

The semantic web aims to provide a universally accessible platform that allows data to be shared and processed by automatic tools as well as by users. It also aims to define and link the data on the web to improve information seeking, discovering, and reuseing across different applications. New languages are developed making more of the information on the web machine-readable. It aims to develop a new generation of technologies and toolkits, as well a new ways to assist the web user.

(1) Semantic web developments include:
   a. Linking of databases (http://www.w3.org/XML)
   b. Sharing content between applications using different XML DTDs or schemas (http://www.w3.org/Xml/Schema), (http://www.w3.org/RDF) (http://www.w3.org/TR/SOAP)
   c. Combination of web services (http://www.w3.org/TR/rdf-schema/) (http://www.w3.org/2001/sw/WedOnt)

This sort of development of the web infrastructure has several merits, including the definition of a shared data model for the design of any query language and the linking of data from many different models. Its limitations however reside in the assumption that the processing of natural languages properties can be dispensed with and that natural language semantics reduces to shallow lexical semantic relations in conjunction with (fragments of) the knowledge of the world. This is illustrated in (2) with SemTag [].

(2) Consider a world in which all documents on the web contained semantic annotations based on TAP. So the sentence: ``The Chicago Bulls announced yesterday that Michael Jordan will...'' would appear as:

```
The <resource ref="http://tap.stanford.edu/
BasketballTeam_Bulls">Chicago Bulls</resource>
announced yesterday that <resource ref=
"http://tap.stanford.edu/AthleteJordan,_Michael">
Michael Jordan</resource> will...''
```

Thus, the annotation:

```
<resource ref="http://tap.stanford.edu/
AthleteJordan,_Michael">Michael Jordan</resource>
```

says that the string ``Michael Jordan'' refers to the resource whose URI is ``http://tap.stanford.edu/AthleteJordan,_Michael.'' It is expected that querying this URI will result in encoded information which provides greater detail about this resource.

.....

* This work is supported in part by funding from the Social Sciences and Humanities Research Council of Canada to the Asymmetry Project, grant number 214-97-0016, as well as by Valorisation-Recherche Québec, grant number 2200-006 attributed to Professor Anna Maria Di Sciullo at the University of Quebec in Montreal.

W3C annotations do not aim to improve real natural language understanding. This is basic however, as efficient web communication.

2. Search engines and their limits

Search engines have been developed to solve the problem of retrieving relevant information within large collections of documents such as the web. The purpose of search engines is to retrieve relevant documents based on the analysis of the queries, the analysis of a set of documents, and a method for determining the relevance of a sub-set of documents with respect to the queries. It is a fact that the performance of current search engines is not optimal and the question that arises is why?

Consider the following simple example illustrating the situation. Yahoo search on the Internet for the query in (3) gives 201,000 results. The first three results, which are ranked as the most relevant, fail to provide the precise information requested. The most relevant documents retrieved include the substantive words of the query in question, such as *government*, *research, funding,* out of the context defined by the query. The most relevant documents retrieved deal **with 'Internet consulting and company development', 'development tools', and 'Web Site development Company', but they do not deal specifically with 'the development of e-commerce in Italy'**.

(3) Query: the development of e-commerce in Italy the funding of research by the government
Results:

1. B2B Tech-Press Releases; B2B web **development**, ecommerce **ebusiness ...**
... was born in **Italy** and has split his time between **Italy** and Kenya. **...** B2B Technologies is a privately held Internet consulting and **development** company specializing **...**
b2btech.com/press-111700.asp cached

2. Magic Software **Italy**- Application **development** tools and **...**
... An investments and business **development** group dealing in **eBusiness** solutions all **...**
www.magicsoftware.com/corporate/executives/default.jsp?branch=it cached | more results from this site

3. Activeunit Web **Development** Company - **E-commerce**
... ActiveUnit can build custom **Ecommerce** applications and **...** PayGate » GZS:PayMaster » NR:**e-commerce** » PayBox » Protx **...** 2003 Activeunit Web Site **Development** Company **...**
www.activeunit.com/**e_commerce**.html cached

In current information retrieval systems, key word search, Boolean analysis, and the use of statistical methods to determine the relevance of documents constitute the standards. The limitations of most current search engines are basically due to the fact that they disregard functional categories (so-called 'stop words'), they relate key words with the concatenation and the Boolean operators, they search for some or a combination of key words and not for linguistic relations. Thus, it comes as no surprise that operating search engines are not optimal.

(4) **Sources of limitations**
-Key word search
-Boolean analysis
-No processing of functional categories
-No recovery of NL relations.

Given the symmetric properties of these operators, the terms of a query can be reordered, and different orderings of the keywords are considered equivalent. The terms of a query can be part of the retrieved documents which may not be entirely relevant or not be relevant at all, such as documents on (5).

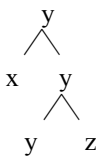
(5) The funding of the government by research

We claim that it is possible to determine with greater precision the information requested by a query by processing the functional words, the so-called stop words, and the asymmetric relations of which they are part. Their articulation into cascades of constituents is determinant for the well-formedness and the interpretation of linguistic expressions.

In Set Theory, an asymmetric relation holds for the ordered pairs in a set if the pair < x, y> is part of that set but not the pair < y, x>. This is the case in (7). The syntactic properties of linguistic expressions are represented by oriented graphs (T), as in (8) and asymmetry holds or not in the set of ordered pairs of T. The relations precede, dominate and sister-contain are asymmetric as if they hold for the ordered pair <x, y>, they do not hold for the pair <y, x>. The relation sister (x, y) is symmetric as the coordinates that are part of the relation can be inversed.

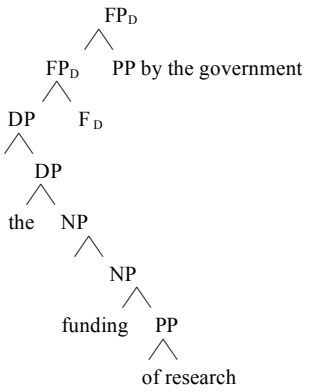
(6) {x, y, z}
(7) {<x, y>, <y, z>, <x, z>}

(8)



The terms of a linguistic expression are articulated on the basis of asymmetric relations, that is, on the basis of oriented dependencies between each pair of terms. This is illustrated in (6), where the constituents' boundaries are identified by the functional categories (Determiner Phrase (DP), headed by the functional category Determiner (FP_D), and Prepositional Phrase (PP), headed by the functional category Preposition (P)). Each nominal term, is in the complement of a functional Head. The first PP *of research* is the complement of the term *funding* and the second PP *by the government* is the adjunct to the whole DP.

(9)



The terms of linguistic expressions are part of asymmetric relations headed by functional heads, such as *of*, *the*, *by*, not by Boolean operators. Thus, information retrieval and extraction systems based on keyword search and Boolean analysis, will bring back wrong results including documents about '*research*' and documents about '*the funding of research',*, as well as documents that will not be relevant at all, including the terms '*government*', '*research'*, and '*funding*', each of these terms interspersed within a document. This introduces what is generally referred to as 'noise' in the information retrieval and extraction process.

Another limitation of most search engines is that they do not take into consideration the asymmetric properties of word-structure. In most cases, the processing of word structure is limited to the normalization of keywords to their uninflected form (blind stemming). Words are formed of roots and affixes (derivational and inflectional) and the relations between them is asymmetric, as demonstrated in Di Sciullo 1999, 2004. The analysis of substantive elements in terms of their internal affix-root relations is absent of most search engines. One consequence of this fact is that

documents that are relevant to a given query, let us say the ones in (10) with respect to the query in (3), will not be retrieved, causing so-called 'silences' to occur in the process.

(10) a. The governmental funds for research
b. Government's funds for research
c. The government funding the researchers

The elimination of affixes by stemming has dramatic consequences for current search engines, as crucial information about the category of a term in a linguistic expression is provided by inflectional affixes, and the semantic type is provided by the derivational affixes. Inflectional affixes, such as number, gender and person, provide core information for anaphoric relations. Thus blind stemming may give rise to wrong results for part-of-speech tagging as well as super-tagging.

Efficient Information Processing is based on the recovery of natural language asymmetries. Together with the identification of functional elements, asymmetric relations contribute to improve the performance of search engines.

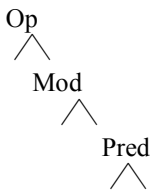
3. Asymmetry theory

Asymmetric relations hold for the objects generated by the grammar, morphological and syntactic. The recovery of the natural language asymmetries constitutes a crucial aspect of the interaction between the grammar (linguistic knowledge) and the performance systems. According to Asymmetry Theory (AT) (Di Sciullo 2004), the Language Faculty is designed to optimally analyze linguistic expressions in terms of asymmetric relations and the external systems are designed to optimally recover natural language asymmetries.

The basic structure of linguistic expressions includes functional elements, operators (NL_{Op}), and modifiers (NL_{Mod}) asymmetrically related to predicates (NL_{Pred}), as schematized below.

(11) a. NL_{Op}: DET, PRON, REFL, PREP, AUX, MOD, COMP, Q, QU
b. NL_{Mod}: AP, ADVP, PP, REL
c. NL_{pred} : NP, VP

(12)



AT is a theory of grammatical relations. It makes predictions about the linear order and the scope relations of the elements of the linguistic expressions. The theory is based on the hypothesis that asymmetric relations, i.e. the precede (13a), dominate (13b), and sister-contain relations (13c), are primitives of the Language Faculty. The operations of the grammar, (14), apply to trees (T), in parallel dimensions of the computational space (CS).

(13) a. [x y] b. x c. x
|
y w x
x y

(14) SHIFT (T_1, T_2), LINK (T), TRANSFER (T), FLIP (T)

The conditions of the grammar, (15), are Economy conditions. The grammatical relations must be legible by the external systems and interpretation is optimal under strict asymmetric relations.

(15) STRICT ASYMEMTRY(□, □), LEGIBILITY (□, □)

As the irreversibility of the terms of a relation is a diagnostic of its asymmetric property. AT correctly predicts that two elements of a linguistic expression cannot be inversed without giving rise to either gibberish or difference in the interpretation. This prediction is borne out as evidenced below with phrasal constituents (16), verb-particle constructions (17), compounds (18), and affixes (19), (20).

(16) a. the funding of research by the government
b. the funding of the government by research
(17) a. He slept over./ He over slept.
b. The vase broke out./ The outbreak of the disease.
c. Dr. No worked over the weekend./ Dr. No overworked his students.
(18) a. The paper-cutter is on the table./ *The cutter-paper is on the table.
b. Bring your walk-man./ *Bring your man-walk.
(19) a. the writer /*er-write of the Book of the Heavens
b. the rewriting/ *ing-write-re of the Book of the Heavens
(20) a. They reentrapped the mouse.
b. *They enretrapped the mouse.

Inversion is possible in structures such as (21)-(23), which relations have been claimed to be symmetric (Den Dikken 1999, Kayne 1994, Zamparelli 1995, Moro 2000). Their symmetric property was questioned in Collins 1997 and Guéron 19XX. Assuming that they include symmetric, they must undergo movement to destroy the symmetry. Assuming they are asymmetric relations, they do not falsify our prediction, as different orders yield a difference in information structure.

(21) a. a picture of the wall is [t the cause of the riot]
b. the cause of the riot is [a picture of the wall t]
(22) a. John bought [books of [t this type]]
b. John bought [this type of [books t]]
(23) a. you are [t kind]
b. it's [kind of [you t]]

Asymmetry determines the linear ordering of the linguistic constituents (precede relation) and their dependencies (dominance and sister-containment). Asymmetry is also crucial for semantic interpretation, including anaphoric relations, allowing anaphors and pronouns to identify a proper antecedent (24), as well as operator-variable relations allowing expressions including operators such as wh-words to bind a variable in their sister-contain domain, (25).

(24) a. the man who wrote [Morphology$_i$ by Itself$_i$]
b. [[the man]$_i$ who wrote Morphology by himself$_i$]
(25) a. Who$_i$ --$_i$ invented electrodynamics?
b. What$_i$ did Einstein discover --$_i$?

Thus, the asymmetry between the parts of the linguistic expressions, that is, relations including functional categories and their dependents, determine their well-formedness and their interpretation.

4. Linguistic infrastructure for the web

We consider the consequences of AT for the overall architecture of the linguistic processor, as well as for the morpho-syntactic analysis of the information conveyed by linguistic expressions.

As existing Information Processing systems are beginning to integrate articulated knowledge of the grammar of natural languages (Copestake and Briscoe 1996, Paziena 1997, Pustejovsky et al. 1997), we show that the availability of asymmetry-based parsing, indexing and search, increases their performance.

The architecture of the linguistic processor in standard Information Processing systems generally includes Stemmers, Part-of-speech (POS) taggers as well as Super-taggers, as schematized below.

(26) **Linear linguistic processor**

Source documents → Stemmer ↔ Lexicon
↓
POS tagger
↓
Super-tagger

A first consequence of our proposal is that the architecture of the linguistic processor may depart from the standard linear ordering of modules, as in (23) and become a dynamic parallel linguistic processing, as in (23'), where the morphological analysis is performed interactively in conjunction with the recovery of the morpho-syntactic asymmetries.

(27) **Dynamic linguistic processor**

Source documents → Asymmetry recovering parser ↔ Lexicon
(queries, texts)
Mophology ↔ syntax
Semantics

Asymmetry-based analysis increases the accuracy of Information Processing. Stemming algorithms are generally ad hoc and offer mediocre results (Porter algorithm (Porter 1980), KIMMO (Karunnen 1983; Antworth 1990). If stemming is performed on the basis of the recovery of morphological asymmetries, including the identification of the categorial head of words, standardization of words is improved. The recovery of the affix-root asymmetries is crucial in the determination of the articulation of words, as a given derivational affix may combine with roots on the basis of their configurational properties (Di Sciullo 1999, 2004). The recovery of the asymmetric structure of derived words improves stemming, as well as it can be used to expand the terms of a query, and thus enhances recall in Information Retrieval and extraction systems. There is evidence to the effect that words are configurations. Asymmetry plays a role in the structuring of derived words. Both derivational and inflectional affixes head words. While an inflectional affix does not determine the category of the word of which it is part, a derivational affix does have this property. This difference follows from the Relativized Head Rule (Di Sciullo and Williams 1987) according to which the category of a word is the category of the rightmost element of that word, which is specified for a categorial feature, ex.: $process_V$-ing_N, $process_V$-or_N-s_{Nl}

Asymmetry is also crucial in part-of-speech (POS) tagging. The category of a word may be predicted from the category of the derivational affix which is the categorial head of that word. Most POS tagging modules are statistically based, or based on distributional repair strategies [6]. In our view, POS tagging is a relational process and not a categorial one. It dynamically interacts with stemming, as inflectional and derivational heads provide categorial information for POS tagging. In our view, POS tagging is based on the asymmetry relating a functional head to its non-head. Thus, it is the asymmetric relation between a functional element and the categories contained in its complement domain, ex.: *to* $_V$***cut*** *an apple* vs. *the* $_A$***cut*** *apple; it* $_V$***contained*** *an*

antecedent vs. *antecedent* $_A$ ***contained*** *deletion.* Categorial deambiguation is based on asymmetric relations, rather than on the choice of one category in a set of possible categories.

Asymmetry is also at play in the identification of the head of syntactic constituents (Super-tagging). This process is not independent from POS tagging and stemming, as a syntactic constituent must have a head. Assuming that constituents are binary branching projections, a syntactic head is asymmetrically related to its non-head. The recovery of asymmetric relations improves syntactic analysis. Misanalysis may arise for example from N/V conversion in English, ex: *how do you monitor costs?*, where taken in isolation both *monitor* and *costs* can be analyzed as parts of N or V projections. Optimal processing can be achieved with the parsing of functional asymmetries.

The recovery/parsing of linguistic asymmetries allows for a reorganization of the architecture of Information Processing systems, including the stemming, the POS tagging , and the Super-tagging modules, see (26) above, as these modules present three sides of the same asymmetry recovering process, see (27) above.

In current operational Information Retrieval systems, both the document analysis and the matching procedure are typically performed with the use of statistical analyses. The role of linguistic, and in particular asymmetry-based grammatical knowledge, contributes to improve the analyses of queries and documents. Consequently, the matching of a query with documents is sharpened rather than being statistically limited. The system achieves greater precision than those based on Boolean relations or other sorts of relations that obviate natural language specific properties. This procedure makes unprecedented use of our knowledge of the grammar of natural languages in Information Retrieval and Extraction.

5. Conclusion

It is generally the case that Information Processing reduces to the processing of the actual substantive words that are part of linguistic expressions. In our view, the information is supported by asymmetric relations headed by functional elements. The consequences of our proposal range over the properties of the architecture of Information Processing systems and the properties of their modules.

Our proposal provides Information Processing systems with the necessary theoretical and linguistic tools to improve their performance. The integration of asymmetry-based parsing in extraction and retrieval systems naturally increases their performance. Information Processing systems integrating natural language asymmetries is a crucial ingredient of Internet Infrastructures ensuring greater efficiency to internet communication.

7. References

E. Antworth, PC-KIMMO: A Two Level Processor for Morphological Analysis, Dallas, TX: Summer Institute of Linguistics, 1990.

A.T. Arampatzis, T. Tsoris and C.H.A. Koster, IRENA: Information Retrieval Engine Based on Natural Language Analysis, RIAO 97 Proceedings, McGill University, 1997.

E. Brill, “Some Advances in Transformation-Based Part of Speech Tagging”, Proceedings of the 12th National Conference on Artificial Intelligence, AAAI 94, 1994

E. Brill,. “A Simple Rule-Based Part of Speech Tagger”, Proceedings of the Third Conference on Applied Natural Language Processing, ACL, 1992.

M. Brody, “Mirror Theory”, Linguistic Analysis 31:1, 2000.

N. Chomsky, Minimalist Inquiries, Ms. MIT, 1998.

N. Chomsky, The Minimalist Program, Cambridge, Mass.: The MIT Press, 1995.

C. Collins, Local Economy, Cambridge, Mass.: The MIT Press, 1997.

A. Copestake and T. Briscoe, “Semi-Productive Polysemy and Sense Extension”, LexicalSemantics. The Problem of Polysemy, Pustejovsky & B. Boguraev (eds.), Oxford University Press, 1996.

S.Derose, “Grammatical Category Disambiguating by Statistical Optimization”,Computational Linguistics 14, 1988.

S. Dill et al. SemTag and Seeker: “Bootstrapping the semantic web vis automated semantic annotation”. WWW2003.

A.M. Di Sciullo. Asymmetry in Morphology, Cambridge, Mass: The MIT Press. Forthcoming.

A. M. Di Sciullo, “Formal Context and Morphological Analysis”, CONTEXT 99, P.Bouquet & al. (eds.), Springler Publishers, 1999b, pp. 105-119.

A.M. Di Sciullo, “The Local Asymmetry Connection”, MIT Papers on Linguistics, Cambridge, Mass: The MIT Press., 1999a.

A.M. Di Sciullo and S. Fong, Morpho-Syntax Parsing, In A.M. Di Sciullo & R. Delmonte (eds.) UG and the External Systems. Amsterdam, John Benjamins. Forthcoming.

A.M. Di Sciullo and E. Williams, On the Definition of Word, Cambridge, Mass.: The MIT Press, 1987.

R. Gaizauskas and A. Roberston, “Coupling Information Retrieval and Information Extraction: A New Text Technology for Gathering Information from the Web”, RIAO 97, Montréal, 1997, pp. 356-370.

K. Hale and J. Keyser, “On the Restricted Nature of Argument Structure”, MIT Papers on Linguistics, Cambridge, Mass: MIT press, 1999

L. Kartunnen, “KIMMO: A General Morphological Processor”, Texas Linguistic Forum 22, 1983.

R. Kayne, The Antisymmetry of Syntax, Cambridge, Mass.: The MIT Press, 1994.

D. Knuth, On the Translation of Languages from Left to Right. Information and Control, pp. 607-639.

C..G. Marken, “Parsing the LOB Corpus”, Association of Computational Linguistics Annual Meeting, 1990.

M.T. Paziena, (ed) Information Extraction. A Multidisciplinary Approach to an Emerging Information Technology, Springler Publishers, 1997.

R. Pohlmann and W. Kraaij, “The Effect of Syntactic Phrase Indexing on Retrieval Performance for Dutch Texts”, RIAO 97 Proceedings, McGill University, 1997.

M.F. Porter, An Algorithm for Suffix Stripping Program, 14.3, 1980.

J. Pustejovsky, B. Boguraev, M. Verhagen, P.Buitelaar, M. Johnston, “Semantic Indexing and Typed Hyperlinking”, Natural Language Processing for the World Wide Web. Papers from the 1977 AAAI Spring Symposium, AAAI Press, 1997.

A. Schiller, “Multilingual Finite-State Noun Phrase Extraction”, Proceedings of the ECAI96 Workshop, 1996.