ARISTA Generative Lexicon for Compound Greek Medical Terms

John Kontos, Ioanna Malagardi, Spyros Fountoukis

Department of Informatics
Athens University of Economics and Business
76 Patission St., 104 34 Athens, Hellas
jpk@aueb.gr, ioanna@ilsp.gr

Abstract

A Generative Lexicon for Compound Greek Medical Terms based on the ARISTA method is proposed in this paper. The concept of a representation independent definition-generating lexicon for compound words is introduced in this paper following the ARISTA method. This concept is used as a basis for developing a generative lexicon of Greek compound medical terminology using the senses of their component words expressed in natural language and not in a formal language. A Prolog program that was implemented for this task is presented that is capable of computing implicit relations between the components words in a sublanguage using linguistic and extra linguistic knowledge. An extra linguistic knowledge base containing knowledge derived from the domain or microcosm of the sublanguage is used for supporting the computation of the implicit relations. The performance of the system was evaluated by generating possible senses of the compound words automatically and judging the correctness of the results by comparing them with definitions given in a medical lexicon expressed in the language of the lexicographer.

1. Introduction

A Generative Lexicon for Compound Greek Medical Terms based on the ARISTA method is proposed in this paper. The method called ARISTA (Kontos, 1980, 1983, 1992, 1998a, 1998b, 1999) which stands for Automatic Representation Independent Syllogistic Text Analysis has been proposed as an alternative to the traditional methods of knowledge extraction from text that rely on the formal representation of text content. These traditional methods have been applied so far in lexical processing projects that rely on the formal representation of the natural language definitions of terms. Characteristic examples of such projects that are based on formal representations are the family of GALEN projects (Rector, et al., 1996, 1998), as well as (Schulz & Romacker, 1998) in the medical domain.

The concept of a representation independent definition-generating lexicon for compound words is introduced in this paper following the ARISTA method. This concept is used as a basis for developing a generative lexicon of Greek compound medical terminology using the senses of their component words expressed in natural language and not in a formal language. The work reported is part of a larger human language technology project of our Research Centre. This project aims at the development of a human language technology system for the knowledge intensive support of a health professional in therapy management. The main knowledge sources of this system are medical lexical resources and textual material concerning drugs necessary for the therapy of infectious diseases.

Compound Greek medical terms used in the English medical sublanguage were analysed into their component words. The possible senses of each compound term were generated using a lexicon defining in natural language the senses of their component words. A Prolog program that was implemented for this task is presented that is capable of computing implicit relations between the components words in a sublanguage using linguistic and extra linguistic knowledge. An extra linguistic knowledge base containing knowledge derived from the domain or microcosm of the sublanguage is used for supporting the computation of the implicit relations. The kinds of relations between nouns that are used are derived from Greek grammar (Malagardi, 1996).

The performance of the system was evaluated by generating possible definitions of the compound words automatically and judging the correctness of the results by comparing them with definitions given in a medical lexicon expressed in the language of the lexicographer. The comparison involved the judgment by health professionals of the acceptability of the definition of the compound word in the cases of wording different than the one given by the lexicographer.

The method proposed could be used to compress the Greek part of a medical lexicon for the space critical applications of systems that will be used with a local interface for the health professionals that will interact with the system remotely. This remote interaction is envisaged to take place using the lightest possible hardware in order to facilitate the use of the system in as many emergency situations as possible.

2. The Computation of Implicit Relations

A Prolog program was implemented and is presented here that is capable of computing implicit relations between the components words in a sublanguage using linguistic and extra linguistic knowledge. An extra linguistic knowledge base containing knowledge derived from the domain or microcosm of the sublanguage is used for supporting the computation of the implicit relations. The kinds of relations between nouns that are used are derived from Greek grammar (Malagardi, 1996). Both taxonomic and meronomic extra linguistic knowledge is used as knowledge base that supports the analysis of the Greek language phrases that are being processed by the system.

Some of the possible implicit relations between Greek nouns such as entity names, property names and process names are presented below in Table 1:

	Noun Combination	Relation	Compound Example	Analysis
1.	Entity + Entity	Locative	Chondrosarcoma	sarcoma *at* chondros
2.	Entity + Entity	Origin	Hemarthrosis	hema *with origin* arthrosis
3.	Entity + Entity	Attributive	Staphylococcus	coccus *appearing like* staphylo
4.	Entity + Entity	Content	Azotemia	azote *inside* ema
5.	Entity + Entity	Content	Bacteremia	bacter *inside* ema
6.	Entity + Entity	Causal	Neurotoxin	toxin *causes damage to* neuro
7.	Relation + Entity	Comparative	Anisocoria	coria *have the relation* aniso
8.	Property + Entity	Attributive	Cardiomegaly	megaly *is an attribute of* cardio
9.	Property + Process	Attributive	Amblyopia	amly *is an attribute of* opia
10.	Entity + Process	Motion object	Aerophagy	aero *is moved by* phagy
11.	Entity + Process	Locative object	Dermatomycosis	mycosis *located at* dermato
12.	Entity + Process	Patient object	Acrocyanosis	acro *is the patient of* cyanosis
13.	Entity + Process	Patient object	Tracheotomy	tracheo *is the patient of* tomy
14.	Entity + Process	Use	Sphygmomanometer	manometer *used for* sphygmos

Table 1: Examples of implicit relations

We describe now how for each example of Table 1 we may compute the implicit relation involved. In example (1) the compound consists of two entity-nouns i.e. “sarcoma” which is a tumor and “chondros” which is a body part. Since we know that tumors are located in body parts it follows that the implicit relation is that of location, where chondros is the location of the sarcoma.

In example (2) the compound consists of two entity-nouns i.e. “hema” which is a fluid and “arthrosis” which is a body part. Since we know that liquids may have as origin body parts it follows that the implicit relation is that of origin, where arthrosis is the origin of hema. Note that the original Greek spelling of “hema” is “haima” or “haema” which means blood.

In example (3) the compound consists of two entity-nouns i.e. “staphylo” which is a fruit and “coccus” which is a microorganism. Since we know that fruits and microorganisms may have only common appearence it follows that the implicit relation is an attributive relation of appearance, where the coccus appears like staphylo.

In example (4) the compound consists of two entity-nouns i.e. “azot” which is a gas and “hema” which is a fluid. Since we know that gases maybe contained in fluids it follows that the implicit relation is that of content, where azot is inside hema or haima as explained above in example (2).

In example (5) the compound consists of two entity-nouns i.e. “bacter” which is a microorganism and “hema” which is a fluid. Since we know that microorganisms maybe contained in fluids it follows that the implicit relation is that of content, where azot is inside hema.

In example (6) the compound consists of two entity-nouns i.e. “neuro” which is a body part and “toxin” which is a harmful substance. Since we know that harmful substances damage body parts it follows that the implicit relation is that of causality, where toxin causes damage to neuro.

In example (7) the compound consists of one adjective denoting a comparative relation i.e. “aniso” and one entity-noun “coria”, which is a body part. Since we know that comparative relations may exist between body parts it follows that the implicit relation is that of comparison.

In example (8) the compound consists of one entity-noun i.e. “cardio”, which is a body part and one adjective (property) i.e. “megaly” which concerns to size. Since we know that body parts have size it follows that the implicit relation is an attributive relation.

In example (9) the compound consists of one adjective (property) that concerns rating i.e. “ambly” and one process-noun i.e. “opia” which is a sensing-process. Since we know that sensing can be rated it follows that the implicit relation is an attributive relation.

In example (10) the compound consists of one entity-noun i.e. “aero” which is a gas (in reality a mixture of gases) and one process-noun i.e. “phagy” which is an ingestion process. Since we know that gases can be moved it follows that the implicit relation is a relation between a process and a moving object.

In example (11) the compound consists of one entity-noun i.e. “dermato” which is a body part and one process-noun i.e. “mycosis” which is a disease. Since we know that diseases are located in body parts it follows that the implicit relation is a locative relation between a disease and a body part.

In example (12) the compound consists of one entity-noun i.e. “acro” which is a body part and one process-noun i.e. “cyanosis” which is a discoloration. Since we know that discolorations can apply to body parts it follows that the implicit relation is a patient relation between the discoloration and the body part.

In example (13) the compound consists of one entity-noun i.e. “tracheo” which is a body part and one process-noun i.e. “tomy” which is a surgical procedure. Since we know that surgical procedures apply to body parts it follows that the implicit relation is a patient relation between the tomy and the body part.

In example (14) the compound consists of one entity-noun i.e. “manometer” which is an instrument and one process-noun i.e. “sphygmo” which is a physiologic process. Since we know that instruments are used for monitoring physiologic processes it follows that the implicit relation is that of use.

In order to be able to perform the computation of implicit relations similar to the ones mentioned above our program must have access to a knowledge base of extra linguistic knowledge derived from the medical domain. Both taxonomic and meronomic extra linguistic knowledge is used as knowledge base that supports the analysis of the Greek language phrases that are being processed by the system. Such a knowledge base will also contain rules of the following kind for the computation of the implicit relations:

1. If tumors are located in body parts it follows that the implicit relation is that of location.
2. If liquids may have as origin body parts it follows that the implicit relation is that of origin.
3. If fruits and microorganisms may have only common appearence it follows that the implicit relation is an attributive relation of appearance.
4. If gases maybe contained in fluids it follows that the implicit relation is that of content.
5. If microorganisms maybe contained in fluids it follows that the implicit relation is that of content.
6. If harmful substances damage body parts it follows that the implicit relation is that of causality.
7. If comparative relations may exist between body parts it follows that the implicit relation is that of comparison.
8. If body parts have size it follows that the implicit relation is an attributive relation.
9. If sensing can be rated it follows that the implicit relation is an attributive relation.
10. If gases can be moved it follows that the implicit relation is a relation between a process and a moving object.
11. If diseases are located in body parts it follows that the implicit relation is a locative relation between a disease and a body part.
12. If discolorations can apply to body parts it follows that the implicit relation is a patient relation between the discoloration and the body part.
13. If surgical procedures apply to body parts it follows that the implicit relation is a patient relation between the procedure and the body part.
14. If instruments are used for monitoring physiologic processes it follows that the implicit relation is that of use.

3. The Computation of the Definitions of Compound Terms

The definitions of all possible senses of each compound term were generated using a lexicon defining in natural language the senses of their component words. The ontology implicit in the definitions is used for choosing the appropriate rule of definition combination. A number of rules specifying the relations allowable between general concepts that group the component words also support this generation. A Prolog program that was implemented for this task is related to the one computing implicit relations between the components words in a sublanguage using linguistic and extra linguistic knowledge. An extra linguistic knowledge base containing knowledge derived from the domain or microcosm of the sublanguage that supports the computation of the implicit relations is used in the program. Some examples of definitions of component words are given below as coded in the program using the predicate “means”:

```
means([acro],[the,extremity_ties]).
means([cyanosis],[a,bluish,discoloration]).
means([aero],[air]).
means([phagy],[eating]).
means([phagy],[swallowing]).
means([brady],[abnormal,slowness]).
means([brady],[sluggineshness]).
means([cardia],[the,heart]).
means([cardia],[the,heart,beat]).
means([kinesia],[movement]).
means([kinesia],[the,physical,and,mental,response_es]).
means([pnea],[the,breathing]).
means([cardio],[the,heart]).
means([genic],[caused,by,a,function]).
means([genic],[caused,by,abnormal,function]).
means([megaly],[large,size]).
means([pathy],[disease]).
means([pathy],[disorder]).
means([cardio],[the,heart]).
means([myo],[the,muscle]).
```

Part of the ontology used for processing these definitions is coded in the program as follows:

```
isa([acro],entity).
isa([aero],entity).
isa([cardio],entity).
isa([cardia],entity).
isa([brady],property).
isa([cyanosis],process).
isa([phagy],process).
isa([cardia],entity).
isa([kinesia],process).
isa([pnea],process).
isa([genic],process).
isa([megaly],property).
isa([myo],entity).
isa([pathy],process).
```

A sample of rules for generating definitions of compound terms is given below.

For two component words:

```
 expl(X,Y,_,_,E) :-
means(X,A,S), means(Y,B,S),
isa(X,property),isa(Y,entity),
conc([of],B,C), conc(A,C,E).
```

and for three component words:

```
expl(X,Y,Z,_,E) :-
means(X,A,S), means(Y,B,S), means(Z,C,S),
isa(X,entity), isa(Y,entity), isa(Z,process),
conc([of],B,B1),conc([of],A,A1),conc(C,B1,D),conc(D,A1,E).
```

Using the predicate expl(A,B,C,D,T), and an output function of the form write(A,"+",B,"+",C,"+",D,"=",T), definitions of the kind illustrated in Table 2 are generated by the program.

["acro"]+["cyanosis"]=["a","bluish","discoloration","of","extremity_ties"]
["aero"]+["phagy"]=["eating","of","air"]
["aero"]+["phagy"]=["swallowing","of","air"]
["cardio"]+["genic"]=["caused","by","a","function","of","the","heart"]
["cardio"]+["genic"]=["caused","by","abnormal","function","of","the","heart"]
["cardio"]+["megaly"]=["large","size","of","the","heart"]
["cardio"]+["pathy"]=["disease","of","the","heart"]
["cardio"]+["pathy"]=["disorder","of","the","heart"]
["cardio"]+["myo"]=["a","muscle","of","the","heart"]
["cardio"]+["pathy"]=["disease","of","the","heart"]
["cardio"]+["pathy"]=["disorder","of","the","heart"]
["brady"]+["cardia"]=["abnormal","slowness","of","the","heart"]
["brady"]+["cardia"]=["abnormal","slowness","of","the","heart","beat"]
["brady"]+["kinesia"]=["abnormal","slowness","of","movement"]
["brady"]+["pnea"]=["abnormal","slowness","of","breathing"]
["brady"]+["kinesia"]=["sluggineshness","of","the","physical","and","mental","response_es"]
["cardio"]+["myo"]+["pathy"]=["disease","of","the","muscle","of","the","heart"]
["cardio"]+["myo"]+["pathy"]=["disorder","of","the","muscle","of","the","heart"]

Table 2: Examples of generated definitions

4. Conclusions

The concept of a representation independent definition-generating lexicon for compound words was introduced in this paper following the ARISTA method. This concept was used and proved adequate as a basis for developing a generative lexicon of Greek compound medical terms using the definitions of the senses of their component words expressed in natural language and not in a formal language. The work reported is part of a larger human language technology project aiming at system for the knowledge intensive support of therapy management. The main knowledge sources of the system are medical lexical and textual resources.

The evaluation of the performance of the system generating possible definitions of compound terms automatically was judged to give satisfactory results. The evaluation was performed by comparing the artificial definitions with those given in a medical lexicon expressed in the language of the lexicographer. The comparison involved the judgment by health professionals who rated well the acceptability of the definition of the compound word in the cases of wording different than the one given by the lexicographer.

The method proposed could be used to compress the Greek part of a medical lexicon for the space critical applications of systems that will be used with a local interface for the health professionals that will interact with the system remotely. This remote interaction is envisaged to take place using the lightest possible hardware in order to facilitate the use of the system in as many emergency situations as possible.

5. References

Kontos, J., (1980). Syntax-Directed Processing of Texts with Action Semantics. *Cybernetica, 23, 2* pp. 157-175.

Kontos, J., (1983). Syntax-Directed Fact Retrieval from Texts with a Micro-Computer. *Proceedings of MELECON '83,* Athens.

Kontos, J., (1992). ARISTA: Knowledge Engineering with Scientific Texts. *Information and Software Technology,* Vol. 34, No 9, pp 611-616.

Kontos, J., I. Malagardi, (1998a). Information and Knowledge Extraction from Medical TextsI. *Health Telematics Education Conference.* Athens.

Kontos, J., I. Malagardi, (1998b). Question Answering and Information Extraction from Texts. *EURISCON '98 Third European Robotics, Intelligent Systems & Control Conference*. Athens. Published in *Conference Procedings "Advances in Intelligent Systems: Concepts, Tools and Applications" (Kluwer).* ch. 11, pp. 121-130.

Kontos, J., I. Malagardi, (1999). I. Information Extraction and Knowledge Acquisition from Texts Using Bilingual Question–Answering. *Journal of Intelligent and Robotic Systems* 26 (2): 103-122, October. Kluwer Academic Publishers.

Malagardi, I., (1996). Determination with computer of the implicit relation between the components of noun phrases in a sublanguage. *Proceedings of the 17th Annual Meeting of the Department of Linguistics. Faculty of Philosophy. Aristotle University of Thessaloniki (in Greek)* pp. 508-520.

Schulz, S., M. Romacker, (1998). A case study of reasoning along part-whole hierarchies in Medicine. In: *KEML'98 - Proc. 8th Workshop on Knowledge Engineering: Methods and Languages*. Karlsruhe, Germany, 21-22 January.

Rector, A., J.E. Rogers, P. Pole, (1996). The GALEN High Level Ontology. *Medical Informatics in Europe (MIE),* Copenhagen.

Rector A, et al., (1998). Practical development of re-usable terminologies: GALEN-IN-USE and the GALEN Organisation. *International Journal of Medical Informatics*. Feb; 48(1-3): 85-101.