IV. Resolution of Lexical Ambiguities in Ophthalmology

M. Coyaud

1. Introduction

The content analysis procedures incorporated into the SMART system are based in part on the words included in documents and search requests, and in part on various dictionary mapping and weighting processes. No systematic attempt is normally made to eliminate all lexical ambiguities; however, the procedures actually used will generally assign large weights to those concepts which are semantically appropriate, while other, semantically inappropriate concepts are assigned much lower weights.

A test is being performed to determine whether a complete automatic text disambiguation procedure would produce substantially better retrieval results than the standard text analysis methods now incorporated into the SMART programs. Specifically, a set of disambiguation rules has been generated for the field of ophthalmology designed to replace each input item by a single, hopefully correct, semantic identifier. These rules are later to be incorporated into a test program operating with an appropriate document collection. The disambiguation rules actually used are described in the present report.

In view of the specialization of ophthalmology in the field of the medical sciences, one would expect that lexical ambiguities having some impact on the precision of the retrieval of documents would be

quite rare. To a certain extent, this assumption proves to be true as far as can be verified using a small corpus of about 600 ophthalmological abstracts. Upon examination, less than thirty index words were found with actual multiple meanings in a small corpus in ophthalmology. As a comparison, more than 100 words selected as descriptors were previously found to be ambiguous in a corpus of 500 abstracts in the field of physiological psychology [1]. The classification used for the dictionary and for the operation of the polysemy rules is provided by the Vision Information Center of Harvard (it is based on the MESH dictionary used at the National Library of Medicine with a number of modifications).

2. Procedures for Devising the Polysemy Rules

Two general approaches can be distinguished: the first, a priori, consists in obtaining by introspection a list of contextual criteria allowing the automatic recognition of the meaning of a word in a given context; the second, a posteriori, consists in reading texts or rather concordances for lists of ambiguous words in given corpuses, so that the actual criteria could be easily picked up from the concordance.

The introspective method contains the risk of being irrealistic. The a posteriori, empirical method has therefore been chosen, but in certain cases, some criteria provided by introspection were listed among the criteria found in the concordances.

A first concordance was obtained from a corpus (later: corpus A) of 300 abstracts; rules were devised for the words actually ambiguous in this corpus, as well as for words which are otherwise known to be ambiguous.

A second concordance was obtained later, from a second corpus (corpus B) of 300 abstracts, in order to verify the rules elaborated from the first concordance.

The original rules (part A) are first examined, followed by the results of the confrontation with the concordances of the control corpus (part C).

A) The Rules Inspired by Corpus A.

a. Exclusion List

Initially, an "exclusion list" was prepared, that is, a list of "words" (= any individual sequence of letters separated by two blank spaces) which are considered as unimportant for the retrieval of documents. This exclusion list (i.e. a kind of "antidictionary") consists of about 2500 words of various types: Grammatical words such as the, a, only, does etc.; general words like high, little, interest; and words which are not pertinent in ophthalmology, like document, information, search etc.

The majority of the ambiguous words of corpus A were included in the exclusion list.

b. Dictionary (DIC)

About 2500 word forms and phrases were included in the dictionary. This list of words kept for indexation consists of the following categories:

i) - nonambiguous words

ii) - phrases

iii) - ambiguous words appearing unambiguously in corpus A.

iv) - ambiguous words appearing with ambiguities in corpus A.

We will examine now the points ii), iii), and iv).

Phrases

A good many phrases were entered in the dictionary (DIC) as separate entries in order to avoid the trouble of generating a polysemy rule. The generation of phrases is the best way of avoiding polysemies; whenever possible, this method was used. Sometimes, it is not sufficient to generate a phrase; then, one also prepares a polysemy rule. For instance, acute refers, in corpus A, to the shape of an angle or to the critical state of a disease. The phrase acute angle was entered in DIC. Many phrases do not include ambiguous words; these phrases were entered because they express a unique concept, like ophthalmic artery; on the other hand, B irradiation, vitamin B, B-adrenergic, B wave are phrases which are always entered as separate DIC entries, since it is obvious that B found alone is ambiguous.

Words Otherwise Ambiguous

A high number (about 100) of words which are ambiguous when found in unspecified contexts boil down to unambiguous entries in the ophthalmological abstracts. For a small number of these words, one can predict that they will never occur with a different meaning than the meaning they usually have in this small subfield of medical science; for instance, for the word cataract, one can predict that it will never be found with the meaning "waterfall" in ophthalmology. Although the words of cataract type are not many, one has to admit that whether they will occur in another sense than that which was selected here is unpredictable. The list of such words is as follows:

abberation (sight)
absence (not mental)
accommodation (sight)
acute (disease, not angle; acute angle is a phrase)
administration (drug)
alignment (with visual axis)
capsule (of eye)
cataract (not in river)
cell (physiological, not photoelectric)
channel (lacrymal, not information)
circulation (of blood, not of documents, cars, etc.)
conduction (phyics)
concentration (chemical; high rate of; not camp)
-te
(NB: in the Medical Subject Headings, the only meaning selected for concentration was as a kind of camp)
cone (eye cell)
contract (straighten)
convergence (optics)
crater (hole in body)
culture (biological)
decentralization (not economic)
density (energy)
deficit (not in budget)
deposition (chemical, physical)
depression (not mental; low pressure)
detachment (retinal, vitreous, of membran, of layer)
diffusion (physical)
digital (= finger; not computer)
disorder (not riot)
discrimination (visual, not racial)
-te
distort (not truth)
diverge (not in opinion)
-nt
dominant (hemisphere in brain)
equator (lens)
erosion (retinal)
exploration (transcranial...)
expose (to light)
-ure
explosive (onset of disease)
field (visual)
gas (as a source of disease)
globes (eye)
heart (in body)
inclusion (particles in eye)
intensity (optics)

invade (microbian)
   -sion (blood)
irradiation (by rays; not physiological)
labyrinth (ear)
nuclear (in cell)
operation (surgical)
operative (surgical)
orange (color)
orbit (of eye, not of a spatial capsule)
orbital (not flight)
organic (chemical)
particle (not physical; small object, charcoal)
plastic (equipment)
pole (not geological nor electrical)
precipitate (physics)
          -ion
race (not run)
radius (of circle; not the bone)
reaction (physiological)
reactive (physiological)
receptor (sensorial)
recipient
red (not communist)
regime (diet)
rehabilitation (medical)
resistance (physiological, not electric, nor politic)
response (physiological reaction)
retardation (mental)
retard (mental)
retraction (iris)
retract (iris)
right (contrary of left)
rod (eye cell)
rotate (eye)
rotation (eye)
separation (of a part of body)
shot (disease)
silver (color)
sinus (anatomical, not geometrical)
solution (liquid)
star (in eye)
stellar (figure on the surface of crystalline lens)
superior (location of muscles)
suspension (in liquid)
sympathetic (not kind)
tear (liquid; not verb)
tension (of tissue, not political nor electrical)
tract (optic, in body; not political)

trial (clinical, therapeutic)
tributary (vein, not river)
triton (a drug, not the animal)
uniform (measure, not cloth)
version (movement)
vessel (in body)
vortex (not eddy)
wall (of vessels in body)
weapon (not therapeutic, but real)
wing (of bone, not of bird nor of theater)

Words ambiguous in Corpus A

Rules were made for the following words:

angle	extract
anterior	fibre
apparatus	front
arm	hand
attack	left
chamber	link
complication	light
conception	paper
contrast	posterior
correction	potential
current	sound
	stress

Here are examples of rules using natural language parameters (the abbreviations for the classes are taken from the Vision Information Center thesaurus):

1. chamber = PϕLNPRE, IN VITRϕ + STϕRED IN + MϕIST ==
   0 * CHAMBER, A13

   this means that chamber, when preceded by the words listed after the code PϕLNPRE, will not be taken as a descriptor; in any other case, it will be taken with the meaning of a part of the eye (anterior, posterior...), class A13.

2. current = PϕLNFϕL, METHϕD + TECHNIQUE == ϕ * CURRENT, H
   this means that when followed by method, or technique, the word current will not be taken as a descriptor in the electrical sense, as in any other case.

3. correction = PϕLN, LENS + SPECTACLE == CORRECTION, EO4
   this means that when words like lens or spectacle appear in the same sentence as the ambiguous correction, then it is assigned a specific meaning in ophthalmology (class EO4 of equipments and therapeutic techniques).

4. fibre(s) = PϕL, AO8 == FIBRE, AO8 * PϕL, AO2 == FIBRE, AO2
   when in the environment of fibre(s) at least one concept (descriptor) appears related to the class of the nervous system (AO8), the meaning is nervous fibre; when the context includes one concept related to the class of muscles (AO2), then it is a muscular fibre.

5. extract (ed+ing) = PϕLIMPRE, D == EXTRACT, D
   when immediately preceded by a concept belonging to the class of chemicals (D), the words extract, extracted, extracting have a chemical meaning.

B) Notes to the Polysemy Rules

a. The parameters used to resolve the ambiguities are defined among the English natural words, and among the concepts symbols.

i) Natural language (English) parameters:

POLN : in the sentence
POLNPRE : among the words preceding the ambiguous word
POLNIMPRE: among the words preceding immediately the ambiguous word
POLNFOL : among the words following the ambiguous word.

ii) Concept parameters:

POL : in the sentence
POLIMPRE: immediately preceding the ambiguous word.

b. The symbols are as follows:

= introduces a rule
== introduces a translation into a concept symbol
, introduces a parameter or the class of a concept
+ introduces a new parameter
new suffix
* introduces a new rule
(after a translation, * means that in any other case, the translation is so and so)
() introduces suffixes

c. The forms of the ambiguous words are not suffixable. For instance, current is the only ambiguous form treated here (electrical current, current method); currents is not ambiguous here.

In certain cases, some suffixed forms are ambiguous too; then, the suffixes are introduced between parentheses. For example: extract (ed + ing).

d. The exclusion lists include words used as parameters in the polysemy rules, for instance: the, one, in, method, is, was, etc. It is necessary to check these words and take them out of the exclusion list when the program processes polysemies.

e. The dictionary takes priority over the exclusion list:

i) Words exist which are put by mistake in the exclusion list.

ii) Words have been put in the exclusion list which are members of dictionary phrases. Example: color naming is in the dictionary but name (with its suffixes) is in the exclusion list.

f. List of Rules for Corpus A

The class symbols refer to the VIC thesaurus (see annex)

anterior = PϕLN, located + segment + part + portion + surface + limit == ANTERIϕR, A

apparatus = PϕL, E == APPARATUS, E * PϕL, A + C + G == APPARATUS, A

arm(s) = PϕL, E == ARM, E

attack(s) = PϕL, C == ATTACK, C

angle(s) = PϕL, A13 + C13 == ANGLE, A

chamber = PϕLNPRE, in vitro + stored in + moist == 0

conception = PϕLN, foetal + uterus == CONCEPTION, G01
= PϕL, G01 + A05 + C06 == CONCEPTION, G01

complication = PϕL, C == CϕMPLICATION, C

contrast = PϕLNPRE, in * PϕLNFϕL, to == 0 * CϕNTRAST, X

hand = PϕLNPRE, on other + on one == 0 * HAND, A

left = PϕLNPRE, are, is, was, were, be, been, being == 0 * LEFT, X

light = PϕLNFϕL, of experiment, result, technique == 0 * LIGHT, H

link(s) = PϕL, D = LINK, D

paper(s) = PϕLNPRE, this + that + his == 0

posterior = PϕL, A == PϕSTERIϕR, A

potential = PϕL, H + G01 == PϕTENTIAL, H

sound = PϕLNPRE, LY == 0 * SϕUND, H (1)

stress = PϕLNFϕL, that + the importance == 0 * STRESS, F

---

(1) Preceded by an adverb, sound is generally an adjective and does not need to be indexed. Ex: this method is neurologically sound.

C) The Control of the Rules by Corpus B

The rules which were controlled are included in part A)

a. Words Otherwise Ambiguous

Some of the words of the list for corpus A were found ambiguous in corpus B:

Precipitation was found in corpus A only in the physical sense: fall of particles; in corpus B, it had the sense of "increase in rapidity"; "the common feature was the precipitation of the attacks by light falling on the patient". In this sentence, there are not less than three ambiguous words. The polysemy of attack and light would have been solved with the available table rules. Precipitation has to be given a rule with parameters: If it has in its environment a word belonging to the class of chemicals, then it should be translated by precipitation, as a chemico-physical process (class H). If not, it is to be translated as haste (in the common field, X).

solution always appeared as a liquid in corpus A. In corpus B, it appeared in such sentences as: "the solution is formulated to provide; solution of problem". In this case, it was not indexed. The suggested rule is to look for words belonging to the chemical class in the environment of solution.

silver was considered as a color in corpus A. In corpus B, it was found as a metal ("silver impregnation technique"). We found no way of solving this polysemy with our simple rules.

Except for these three cases, the list of "words otherwise ambiguous" but supposed to be unambiguous in ophthalmology was confirmed by an examination of corpus B.

b. Verification of the Rules with Parameters

Some parameters had to be added to the rules shown in part A. In corpus B, the rules for contrast, correction, anterior, current and potential proved insufficient. The following ameliorations are proposed:

contrast: add by as a parameter: by contrast is not to be indexed.

correction: did not necessarily refer to the spectacles, but to surgery too ("surgical steps for correction of these complications; operation for correction of dystopia").

current: add the parameter criteria.

anterior: add the parameter region, structure.

potential: the parameters were insufficient. The following contexts suggest new parameters: "evoked p., photoreceptor p., the p. appears negative, transcorneal p., absence of any p., ERG p.". In other contexts, potential was not to be indexed (meaning "possible"); "potential pathogens, the true nature and p. severity of rubella, there is a p. histidine pump mechanism".

c. Polysemies Intractable by Simple Means

Some abbreviations are ambiguous which seem very difficult to solve when one cannot generate a phrase. We have already seen, the example B (part A). We have found Cu meaning "candle unit" instead of "copper". Besides silver, here are some polysemies unsolvable by the means devised in this experiment:

net (lattice, pure): "the net effect of the corticofugal influence".

rose (noun, verb): "staining of cornea by rose bengal dye; iodate levels of the fluids rose after".

rest (lay, remain): "the curvature of the cornea on which it is to rest, an eccentric rest position; the rest of the periphery".

3. Conclusion

In a first experiment of polysemy, resolutions made in the frame of a research on the SYNTOL Information Retrieval System [1], almost all the rules were defined with reference to the classifications; in other words, these rules referred to the contexts of concepts (i.e. descriptors) and not to the natural language contexts. These rules worked well in about nine cases out of ten (100 rules were applied 432 times). For improving these results, it was suggested to devise rules using certain grammatical data [2], or rules using contextual parameters of the natural language. This second proposal, obviously simpler, was applied in the present experiment, on about 600 abstracts of ophthalmology; the rules seemed to work reasonably well, at least for this small corpus.

The remaining and essential question is that of the degree of generality of these rules and of their applicability to other corpuses in ophthalmology or other fields.

a. Applicability in Ophthalmology

The only answer to this question is empirical; one has to make further concordances in the same field, for a selection list of ambiguous words, in order to obtain a more reliable assessment of the general validity of our rules.

b. Applicability in Other Fields

The answer is very probably no. One can see that the list of critical words in ophthalmology and in other subfields of medical science, (as epilepsy) or in related fields (like physiological psychology) are approximately the same; but the preferred senses in each of these fields,

and consequently, the polysemy rules are often different. Consequently, the rules presented here for ophthalmology are probably not applicable to other medical subfields.

References

[1] M. Coyaud and N. Siot-Decauville, L'Analyse Automatique des Documents, Section 5 and Annex, Paris, Mouton, 1967.

[2] A. Borillo and J. Virbel, Etudes sur L'Indexation Automatique, Centre National de la Recherche Scientifique, Centre d'Analyse Documentaire, Marseille, France, 1966.

Annex I

Outline of the Thesaurus of VIC

A. Anatomy

- A01 Parts of body
- 2 Muscle, Skeleton
- 3 Digestion
- 4 Respiration
- 5 Urogenital
- 6 Endocrine
- 7 Cardiovasc.
- 8 Nerves
- 9 Sense (except eye)
- 10 Tissues, Embryonic
- 11 Cells, Cells constitutents
- 12 Body Fluids
- 13 Eye

B. Biology

- B.01 Invertebrates
- 2 Vertebrates
- 3 Bacteria
- 4 Viruses
- 5 Fungi

C. Diseases

- C.01 Infectious
- 2 Neoplasms, Cysts, Polyps
- 3 Muscular, Skeleton
- 4 Digestion
- 5 Respiration
- 6 Urogenital
- 7 Endocrine
- 8 Cardiovascular
- 9 Uremic, Lymphatic
- 10 Nervous
- 11 Sense (except eye)
- 12 Intertegumentary
- 13 Eye
- 14 Injury, Poisoning, Shock
- 15 General Disease, Pathology
- 16 Nutrition, Metabolism

D. Drugs, Chemicals

E. E.01 Diagnostic Techniques and Measurement

- 2 Therapeutic
- 3 Anesthetic
- 4 Surgery (eye)
- 5 Lab Techniques - Equipment
- 6 Diagnostic Techniques - Equipment
- 7 General Surgery Techniques - Equipment

F. Psychology

G.01 Biological Functions

H. Physics

(NB) We added a class X for the general words.