A Qualitative Comparison of Scientific and Journalistic Texts from the Perspective of Extracting Definitions

Igal Gabbay
Documents and Linguistic
Technology Group
Department of Computer Science
University of Limerick
Limerick, Ireland
Igal.Gabbay@ul.ie

Richard F. E. Sutcliffe
Documents and Linguistic
Technology Group
Department of Computer Science
University of Limerick
Limerick, Ireland
Richard.Sutcliffe@ul.ie

Abstract

In this paper we highlight a selection of features of scientific text which distinguish it from news stories. We argue that features such as structure, selective use of past tense, voice and stylistic conventions can affect question answering in the scientific domain. We demonstrate this through qualitative observations made while working on retrieving definitions to terms related to salmon fish.

1 Introduction

An information retrieval system informs on the existence (or non-existence) and whereabouts of documents relating to the request of a user (Lancaster, 1968). On the other hand, a question answering (QA) system allows a user to ask a question in natural language and receive a concise answer, possibly with a validating context (Hirschman and Gaizauskas, 2001).

Questions asking about definitions of terms (i.e., 'What is X?') occur frequently in the query logs of search engines (Voorhees, 2003). However, due to their complexity, recent work in the field of question answering has largely neglected them and concentrated instead on answering factoid questions for which the answer is a single word or short phrase (Blair-Goldensohn et al., 2003). Much of this work has been motivated by the question answering track of the Text REtrieval Conference (TREC), which evaluates systems by providing them with a common challenge.

In a recent project inspired by our experiences in TREC (Sutcliffe et al., 2003), a system was built for extracting definitions of technical terms from scientific texts. The topic was salmon fish biology, a very different one from that of news articles. What, then, is the effect of domain on the applicability of QA? In this paper we attempt to answer this question, focusing on definitions and drawing on our findings from previous projects.

The rest of the paper is structured as follows: First, we review recent related work. Second, we summarise the objectives, methods and findings of the SOK-I QA project, named after the sockeye salmon. Third, we compare the characteristics of scientific text with those of newspaper articles illustrating our points with examples from our SOK-I collection as well from the New York Times, CLEF 1994 Los Angeles Times collection and AQUAINT corpus. Fourth, we discuss the implications that these have for definitional QA. Finally, we draw conclusions from the study.

2 Recent Related Work

Zweigenbaum (2003) describes biomedicine as a specialised domain and argues that it is not necessarily simpler than an open domain as is sometimes assumed. He identifies the following characteristics:

- A highly specialised language for both queries and articles;
- A potential difference in technical level between user questions and target documents;
- A problem concerning the variable (and possibly unknown) reliability of source documents and hence that of answers drawn from them;
- A potential for using a taxonomy of general clinical questions to route queries to appropriate knowledge resources.

The gap in technical level between non-expert users and target documents is addressed by Klavans and Muresan (2001). Their system, DEFINDER, mines consumer-oriented full text medical articles for terms and their definitions. The usefulness and readability of the definitions retrieved by DEFINDER were both rated by non-experts as being significantly higher than those of online dictionaries. However, Klavans and Muresan do not focus specifically on the characteristics of the source documents in their domain.

The view of Teufel and Moens (2002) that summarization of scientific articles requires a different approach from the one used in summarization of news articles may perhaps apply to QA. The innovation of their work is in defining principles for content selection specifically for scientific articles. As an example they observe that information fusion (the comparison of results from different sources to eliminate mis-information and minimize the loss of data caused by unexpected phrasing) will be inefficient when summarizing scientific articles, because new ideas are usually the main focus of scientific writing, whereas in the news domain events are frequently repeated over a short time.

The lack of redundancy as a feature of technical domains is also mentioned by Mollá et al. (2003). They argue that because of this and the limited amount of text, data-intensive approaches, which are often used in TREC, do not work well in technical domains. Instead, intensive NLP techniques are required. They also mention formal writing and the use of technical terms not defined in standard lexicons as additional features.

3 Answering Definition Questions Related to Salmon (the SOK-I Project)

Many of the observations in this paper are based on a recent study concerned with answering definition related to salmon (Gabbay, 2004). While a full treatment of the work falls outside the scope of this paper, we summarise the key points here. The objectives of the project were:

- To test the effectiveness of lexical patterns without deep linguistic knowledge in capturing definitions in scientific papers;
- To discover simple features which indicate sentences containing definitions;
- To study the stylistic characteristics of definitions retrieved from scientific text.

We chose the terminology-rich field of salmon fish biology as the research domain. A collection of 1,000 scientific articles (Science Direct, 2003) matching the keyword ‘salmon’ was used as the source of definitions. Most of the documents were in agricultural and biological sciences. Each sentence in the articles was indexed as a separate document.

A system was then developed which could take as input a term (e.g. ‘smolt’) and carry out the following steps:

1. Retrieve all sentences in the collection containing the term;
2. Extract any portions of these which matched a collection of syntactic patterns.

The patterns used were similar to the ones used by Hearst (1992), Joho and Sanderson (2000) and Liu et al. (2003) to retrieve hyponyms from an encyclopedia, descriptive phrases from news articles and definitions from Web pages, respectively.

To evaluate the system four test collections of terms were used: 42 terms which were suggested by salmon researchers, and three collections containing 3,920, 2,000 and 1,120 terms respectively. The latter were extracted from a database on the Web called FishBase (2003). For each collection, the output corresponding to a term was inspected manually and each phrase matching a pattern was judged to be either Vital, Okay, Uncertain or Wrong.

While a complete discussion of the results and methods used to obtain them can be found in Gabbay (2004), the main quantitative finding of the project was that techniques adopted could achieve a Recall of up to 60%.

Drawing from our experiences in SOK-I and TREC, we turn in the next section to some specific observations regarding differences between salmon biology texts and newspaper articles.

4 Scientific and Journalistic Texts Compared

4.1 Outline

From our QA studies in the salmon biology field as well as experiences with news articles in TREC and CLEF, many interesting differences between these areas have come to light which we summarise here. The comparison is divided into six features: structure, tense, voice, references, terminology and style.

4.2 Structure

Scientific articles normally follow the structure known as IMRAD (Introduction, Methods, Results, and Discussion). This is the most common organisation of scientific papers that report original research (Day, 1998). For example, the guidelines to authors submitting papers to the journal Aquaculture (Elsevier Author Guide, 2003) specify the following required sections: Abstract, Keywords, Introduction, Methods and Materials, Results, Discussion, Conclusion, Acknowledgments and References.

The structure of a news story is often described as an inverted pyramid, with the most essential information at the top (Wikipedia, 2004). The most important element is called the *lead* and is comparable to the abstract of scientific articles but limited to one or two sentences (leads are often absent in longer feature articles).

The introduction of a scientific paper on the other hand often begins with general statements about the significance of the topic and its history in the field; the 'news' is generally given later (Teufel and Moens, 2002).

4.3 Tense

In scientific writing it is customary to use past tense when reporting original work and present tense when describing established knowledge (Day, 1998). For example, the following sentence reports an accepted fact:

> 'The idea behind using short-term temperature manipulations to mark juvenile fish otoliths is to alter the appearance of D- and L-zones in one or more increments to produce an obvious pattern of events.' (SD-1)

Contrast this with the sentence

> 'Otoliths (sagittal otoliths) were taken from each fish in the total sample or a subsample of the total catch.' (SD-2)

which describes a technique used specifically in the reported study. Therefore, it is reasonable to expect that verbs in the past tense will be concentrated in the Methods and Results sections.

The past tense seems to dominate journalistic writing. In news reporting the past tense is considered slower, whereas the present tense is used for dramatic effect (Evans, 1972). The following excerpt gives a sense of urgency due to the use of the present progressive:

'Pacific salmon contaminated by industrial pollutants in the ocean are carrying the chemicals to Alaska's lakes…' (NYT-1)

4.4 Voice

The passive voice is a major stylistic feature of scientific discourse where according to Ding (1998) it represents the world in terms of objects, things and materials. Therefore, grammatical subjects are more likely to refer to inanimate objects than to humans.

Journalistic prose generally uses the active voice which is thought to assist in reading comprehension but also reflects the focus of news reporting on people and organizations (and indeed 80% of the definition questions in TREC were about a person or an organisation). For example, compare the first two sentences of a report appearing in the Brief Communication section of the journal Nature to the lead of the same report as it was printed in popularized form in the New York Times:

> 'Pollutants are widely distributed by the atmosphere and the oceans. Contaminants can also be transported by salmon and amplified through the food chain.' (NAT)

> 'Pacific salmon contaminated by industrial pollutants in the ocean are carrying the chemicals to Alaska's lakes, where they may affect people and wildlife…' (NYT-1)

In the first excerpt the subject is the contaminants being transported by the salmon (passive), whereas in the second the subject is the salmon carrying them (active).

4.5 Citations

Previously published work is cited frequently in scientific text using a consistent format such as the Harvard author-year citation style which is being used in this paper. Most of the citations are silent (i.e., both the name(s) and the date are enclosed in brackets) and often appear at the end of sentences.

In the news domain, sources are often quoted directly. If the source is another publication, it is mentioned but rarely referenced in a detailed format with volume, issue, page numbers etc.

For example, the author of the study which was published in Nature is quoted directly:

> '"They die in such huge numbers that it almost looks like you can walk across the lakes", an author of the of the study Dr. Jules Blais, said'. (NYT-1)

People can also be quoted indirectly by reported speech as in the following example,

> 'The salmon act as biological pumps, Dr. Blais said…' (NYT-1)

4.6 Terminology

Specialised terms abound in scientific writing and constitute a jargon. Such terms do not usually appear in news stories. For example, in the entire TREC AQUAINT collection the term 'smolt' appears eight times but more than 1,300 times in the SOK-I collection we created for our project. The term 'smoltification' which appears almost 600 times in SOK-I is missing entirely from AQUAINT.

Journalistic prose relies much less on jargon. Journalists tend to favour short common words over long infrequent ones. Compare the vocabulary of Nature:

> 'Here we show that groups of migrating sockeye salmon (*Oncorhynchus nerka*) can act as bulk-transport vectors of persistent industrial pollutants known as polychlorinated biphenyls (PCBs), which they assimilate from the ocean and then convey over vast distances back to their natal spawning lakes. After spawning, the fish die in their thousands - delivering their toxic cargo to the lake sediment and increasing its PCB content by more than sevenfold when the density of returning salmon is high.' (NAT)

to the same story in the New York Times:

> 'After spending most of their lives in the ocean, where they absorb widespread industrial chemicals like PCB's, sockeye salmon flock to Alaska's interior lakes in huge numbers to spawn and then die. Each salmon accumulates just a small quantity of PCB's. But when the fish die together in the thousands, their decaying carcasses produce a sevenfold increase in the PCB concentrations of the spawning lakes, the study found.' (NYT-1)

Note, for example, that the abbreviation 'PCB' is never expanded in the New York Times report. Presumably, the precise chemical name is of little interest to the average reader of the Times, whereas in scientific text there is a need to avoid any technical ambiguity. The Nature report also uses the more technical terms 'vectors' 'assimilate' and 'sediment'.

4.7 Style

Apart from a particular citation style, which is a dominant feature of scientific text, entities such as species or chemical compounds are usually written according to standard nomenclature and format. For example, the common name of an animal species is normally followed by the binomial scientific name in italics and often bracketed:

> 'Here we show that groups of migrating sockeye salmon (*Oncorhynchus nerka*) can act…' (NAT)

News stories usually only use the common name of a species (e.g. sockeye salmon).

In the next section we will see how such features affect definitional QA.

5 Implications for Definitional QA

5.1 Structure

Blair-Goldensohn, McKeown and Schlaikjer (2003) and Joho and Sanderson (2000) who worked in the news domain observed that definitions are likely to be found nearer the beginning of the document than its end. They relied on relative and absolute sentence position as a feature indicating the presence of definitions. However, our observations suggest that at least in the SOK-I collection, sentence position (either relative or absolute) is not a good indicator of text containing definitions. This might be the result of the structured organisation of scientific papers, where each section is more self-contained than paragraphs are in news reports. We expected to find most of the definitions in the Introduction but other sections yielded many definitions. Early in the project we considered discarding the References section during the document pre-processing stage but later discovered it can contain definitions such as:

> 'Canthaxanthin: a pigmenter for salmonids' (SD-3)

However, definitions from different sections of the paper may differ in nature and style. For instance, definitions extracted from the Methods are more technical:

> 'Dry matter eaten was defined as dry matter waste feed collected divided by recovery percentage, subtracted from the dry matter fed.' (SD-4)

It is worth exploring whether certain types of terms are more likely to be defined in particular sections. A similar approach was suggested by Shah at al. (2003) for extracting keywords from full-text papers in genetics.

5.2 Tense

Since the present tense is often used to state established knowledge, we expected that lexical patterns in the present tense would be more likely to match definitions to terms. We observed that many of the wrong answers in our output matched the past tense version of the copular pattern (**TERM was/were DEFINITION**). Sometimes, however, actions performed on or by the term can elucidate it. This is especially common in the Methods section of papers. For example, the term 'Secchi disc' is defined in FishBase as:

> 'A 20 cm diameter disc marked in 2 black and 2 white opposing quadrants, lowered into the water. The average of the depth at which it disappears from sight and the depth at which it reappears when lowered and raised in the water column is the Secchi disc reading, a measure of transparency.'

We retrieved the following answer which was judged as Okay:

> 'Secchi disc was used to measure water visibility (m of visibility) at 1400h…' (SD-5)

5.3 Voice

Certain lexical patterns for definitions are in passive voice. For example the pattern **DEFINITION is termed TERM** matched the following sentence in the SOK-I collection:

> 'The best-known physical damage caused by aggression is inflicted on the fins and is termed fin damage, fin erosion or fin rot.' (SD-6)

On the other hand definitions to technical terms in news stories are more likely to be attached to their definers—experts such as 'biologists' in the following example:

> 'human illness from the virus will probably remain rare since humans are likely to remain what **biologists call** ``dead-end hosts": they can be infected, but their immune systems almost always prevent the virus from multiplying enough to be passed back to mosquitoes and then to other hosts.' (NYT-2)

5.4 Citations

One of the most common definition patterns is a term followed by its definition in brackets:

> 'Grilse (fish maturing after 1.5 years in sea water)' (SD-7)

In our first experiment we observed that the pattern falsely matched citations, and references to figures and tables as in the following case:

> 'redd (Fleming, 1998)' (SD-8)

These were eliminated by creating a list of stopwords which are typical to bracketed references (e.g., 'et al.', 'fig.', years).

Sometimes we encountered names of cited authors which matched a term to be defined or part of it (e.g. Fry, Fish). In the future these names need to be disambiguated.

5.5 Terminology

Definitions in scientific text are generally more technical and precise than in the news domain. For example, in SOK-I we matched the following definition of smolt:

> 'In Atlantic salmon culture, smolt is usually defined as a juvenile salmon that is able to survive and grow normally in sea water.' (SD-9)

In a newspaper we may find 'smolts' defined as in the following sentence:

> 'Young, six-inch-long first-year salmon, called by the old Anglo-Saxon name of smolts, migrate to two main oceanic feeding areas from their home streams in New England…' (NYT-3)

In the last definition the focus was on the word 'smolt' which may be foreign to many newspaper readers. On the other hand, the readers of scientific papers on salmon biology are probably familiar with the term but may need to know its exact usage.

Scientific names of species are taxonomically informative to biologists but would normally mean little to a non-expert. For instance, in scientific text 'steelhead trout' would be followed by its scientific name *Oncorhynchus mykiss* which tells the informed reader it is a species of the same genus to which other pacific salmons belong. In a news articles, we found the following sentences:

> 'But in this case, the endangered animal is the steelhead trout, a relative of the salmon…' (LA-1)

> 'Copper River king salmon, magnificent sea beasts as big and fleshy as Chinese temple dogs, had been running…' (LA-2)

Often definitions of species and other terms will just burden the readers of a newspaper and therefore are unnecessary. For example, unlike biologists, they do not require an exact definition of 'salt water' which specifies the concentration of salt or of 'colour' in the context of salmon meat quality.

Sometimes definitions retrieved from scientific text were found to contain terms which would have to be defined in a news article. For example 'smolt' can be defined in terms of degree days—the product of the daily water temperature, multiplied by the number of days it takes the salmon to reach the smolt stage.

Even though the papers in the SOK-I collection seemed to target a homogenous audience, it was possible to find definitions which are suitable for different levels of expertise. For instance, the system retrieved the chemical name

> '(3,3'-dihydroxy-,-carotene-4,4'-dione)' (SD-10)

in response to the query 'astaxanthin'. Such an answer, although incomplete, could satisfy an expert in biochemistry. Another answer was:

> 'Astaxanthin is an approved colour additive in the feed of salmonids' (SD-11)

The first definition was found in a biochemistry paper on the digestability and accumulation of astaxanthin, whereas the second one was extracted from a fishery research paper which discusses potential issues for human health and safety from net-pen salmon farming. The readers of the second paper may be experts on fish biology but not necessarily on chemicals, food safety or even salmon farming, whereas the first paper is more limited to a single discipline.

5.6 Style

The standardised forms of species and chemical names in scientific text lend themselves to information extraction techniques which would not be effective in the news domain. Templates could be created for certain categories of biological terms. For example, for the category Species we can fill the slots for the scientific name, taxonomic family or order, distribution, life cycle, synonym, and threats to the species; In our experiments the pattern **TERM (DEFINTION)** was effective in recognising the scientific name when the query term was the common name of a species.

6 Conclusions

In this paper, we demonstrated how scientific and journalistic texts differ in structure, tense, voice, references, terminology and style. Our observations are based on a project in which we retrieved definitions to terms in the salmon fish domain. The above features could be exploited specifically in scientific QA. Features such as voice may play a more significant role in QA systems which employ deeper NLP techniques than the simple patterns we used. The uniform structure of scientific documents may allow us to typify definitions in each section before combining them to suit the need of users. Further analysis of the news domain may perhaps yield more observations which will also contribute to current mainstream open-domain QA research as seen in TREC and CLEF.

7. Sources of Cited Examples

LA-1: CLEF 03 LA Times LA120894-0019

LA-2: CLEF 03 LA Times LA070794-0021

NAT: Nature, 425(6955), 255.

NYT-1: http://www.nytimes.com/2003/09/23/science/23SALM.html?ex=1079499600&en=1083ff4683d95e8e&ei=5070

NYT-2: AQUAINT NYT20000807.0291

NYT-3: AQUAINT NYT19990913.0215

SD-1: http://www.sciencedirect.com/science/article/B6T6N-3XNJYSC-P/2/704eaa76fae2ceb6b79ec11d844a44dd

SD-2: http://www.sciencedirect.com/science/article/B6T6N-409630G-V/2/9bc703e5948960159743f99269998fb6

SD-3: http://www.sciencedirect.com/science/article/B6T4D-428FK2P-C/2/d9e2c93377b34beb5ecc47165a4b1098

SD-4: http://www.sciencedirect.com/science/article/B6T4D-460WH4M-1/2/fba8df4de8a057a606cc582a60046c09

SD-5: http://www.sciencedirect.com/science/article/B6T4C-43X1B91-4/2/b7b7058db2c954eaf9d72b7bf2b5d141

SD-6: http://www.sciencedirect.com/science/article/B6T4D-3YXJYY2-9/2/e3ddd1ccfbbece503a0ae26304a4b443

SD-7:

http://www.sciencedirect.com/science/article/B6T4D-3WN6GV4-/2/13a6aa37050ce68845d85c8eb111a82b

SD-8:

http://www.sciencedirect.com/science/article/B6T6N-472BJBX-4/2/054a7ff897821495aed30fe697c3b1c7

SD-9:

http://www.sciencedirect.com/science/article/B6T4D-40FG8N8-G/2/459f344b039746dc9dff2a3ca1f17679

SD-10:

http://www.sciencedirect.com/science/article/B6T2R-41JM957-K/2/5317d1c1daefee0ddadbc86a31288eb2

SD-11:

http://www.sciencedirect.com/science/article/B6T6N-4846K7G-2/2/1c8d6922218aabc83ad653b376d39ed9

References

Blair-Goldensohn, S., McKeown, K. R. and Schlaikjer, A. H. (2003). Retrieved November 30, 2003. http://trec.nist.gov/act_part/ t12_notebook/papers/columbiau.qa.pdf

Day, R. A. (1998) How to write & publish a scientific paper, Cambridge University Press, Cambridge.

Ding, D. (1998) In Essays in the study of scientific discourse : methods, practice, and pedagogy (Ed, Battalio, J. T.) Albex Publishing, Stamford, CT, pp. 117-138.

Elsevier Author Guide (2003). Retrieved December 20, 2003. http://authors.elsevier.com/ GuideForAuthors.html

Evans, H. (1972) Editing and design : a five-volume manual of English, typography and layout--Book 1 : Newsman's English, Heineman, London.

FishBase (2003). http:\\www.fishbase.org

Gabbay, I. 2004. Retrieving Definitions from Scientific Text in the Salmon Fish Domain by Lexical Pattern Matching. MA thesis in Technical Communication, University of Limerick, Limerick, Ireland.

Hearst, M. A. (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora". In Proceedings of the 14th International Conference on Computational Linguistics (COLING-'92). Nantes, France, ed. 539-545.

Hirschman, L. and Gaizauskas, R. (2001) "Natural Language Question Answering: The View from Here". Journal of Natural Language Engineering, 7(4), 325–342.

Joho, H. and Sanderson, M. (2000). "Retrieving Descriptive Phrases from Large Amounts of Free Text". In Proceedings of the ninth international conference on Information and knowledge management (CIKM). McLean, VA, ed. 180-186.

Klavans, J. and Muresan, S. (2001). "Evaluation of DEFINDER: a system to mine definitions from consumer-oriented medical text". In ACM/IEEE Joint Conference on Digital Libraries, JCDL 2001. Roanoke, Virginia. 201-202.

Lancaster, F. W. (1968) Information Retrieval Systems: Characteristics, Testing and Evaluation, Wiley, New York.

Liu, B., Wee, C. and Ng, H. T. (2003). "Mining topic-specific concepts and definitions on the web". In Proceedings of the twelfth international conference on World Wide Web. Budapest, Hungary. 251 - 260.

Mollá, D., Schwitter, R., Rinaldi, F., Dowdall, J. and Hess, M. (2003). "NLP for Answer Extraction in Technical Domains". In Workshop on Natural Language Processing for Question Answering, EACL 2003. Budapest, de Rijke, M. and Webber, B., eds. 5-11.

Shah, P., Perez-Iratxeta, C., Bork, P. and Andrade, M. (2003) "Information extraction from full text scientific articles: Where are the keywords?" BMC Bioinformatics, 4(1), 20.

Sutcliffe, R. F. E., Gabbay, I., Mulcahy, M. and White, K. (2003). Retrieved November 2003 http://trec.nist.gov/ act_part/t12_notebook/papers/ulimerick.qa.pdf

Teufel, S. and Moens, M. (2002) "Summarizing scientific articles: experiments with relevance and rhetorical status". Computational Linguistics, 28(4), 409-445.

Voorhees, E. (2003). Retrieved November 30, 2003. http://trec.nist.gov/act_part/t12_notebook/t12_notebook.html

Wikipedia (2004). Retrieved March 10, 2004. http://en.wikipedia.org/wiki/News_style

Zweigenbaum, P. (2003). "Question answering in biomedicine". In Workshop on Natural Language Processing for Question Answering, EACL 2003. Budapest, de Rijke, M. and Webber, B., eds. 1-4.