Inf Retrieval (2007) 10:393–414
DOI 10.1007/s10791-007-9029-5

Syntactic sentence compression in the biomedical domain: facilitating access to related articles

Jimmy Lin · W. John Wilbur

Received: 12 February 2007 / Accepted: 27 June 2007 / Published online: 21 August 2007

Abstract We explore a syntactic approach to sentence compression in the biomedical domain, grounded in the context of result presentation for related article search in the PubMed search engine. By automatically trimming inessential fragments of article titles, a system can effectively display more results in the same amount of space. Our implemented prototype operates by applying a sequence of syntactic trimming rules over the parse trees of article titles. Two separate studies were conducted using a corpus of manually compressed examples from MEDLINE: an automatic evaluation using BLEU and a summative evaluation involving human assessors. Experiments show that a syntactic approach to sentence compression is effective in the biomedical domain and that the presentation of compressed article titles supports accurate "interest judgments", decisions by users as to whether an article is worth examining in more detail.

Keywords Sentence compression · Extrinsic evaluation · PubMed · MEDLINE · Genomics IR

1 Introduction

Sentence compression has previously been identified as a key component in document summarization. Indeed, the ability to convey the substance of a piece of text, but in a smaller amount of space, is one hallmark of a good summarization system. Although both syntactic and statistical approaches have been employed to tackle this problem, previous attempts largely focus on newswire text. This work differs from previous studies of sentence compression in two important ways: First, we explore an application

J. Lin (✉)
College of Information Studies, University of Maryland, College Park, MD, USA
e-mail: jimmylin@umd.edu

W. J. Wilbur
National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, USA
e-mail: wilbur@ncbi.nlm.nih.gov

of syntactic trimming techniques in the biomedical domain. Other than brief mentions by Ruch et al. (2003) and Lu et al. (2006) in the context of GeneRIF extraction, this work represents the first systematic attempt, to our knowledge, at tackling sentence compression in this highly-specialized domain. Second, we couch sentence compression within the context of result presentation in an information retrieval task. This framing of the problem provides an extrinsic, task-based evaluation grounded in real-world user scenarios.

We present a sentence compression algorithm for article titles from MEDLINE based on syntactic trimming rules that operate over parse trees. This approach was adopted due to its proven effectiveness in previous summarization tasks and the paucity of training data in the biomedical domain. Intrinsic evaluations show that trimmed titles generated by our system are competitive with manually-compressed gold standards, both in terms of automatic metrics (BLEU) and human judgments of content and fluency. In addition, we evaluated the ability of our compressed titles to facilitate "interest judgments"—the decision by a user regarding whether or not an article is worth examining in response to an information need. We found little difference between original and automatically compressed titles, indicating that from a task viewpoint, our system can effectively support users' decisions while reducing the amount of text they must read.

This article is organized as follows: We begin by describing the task model that underlies our explorations in Sect. "Motivation". Previous work is reviewed in Sect. " Related work". Efforts to develop appropriate data resources are outlined in Sect. " Resource development". The syntactic trimming algorithm is detailed in Sect. "A syntactic approach to compression". Evaluation is broken up into two sections: Sect. "Automatic evaluation" covers automatic evaluations, while Sect. "Manual evaluation" covers manual evaluations. We discuss the significance of our work in Sect. "Discussion" and future plans in Sect. "Future work" before concluding.

2 Motivation

Our work is situated in the context of the PubMed search engine,[1] a freely-accessible gateway to the MEDLINE bibliographic database maintained by the US National Library of Medicine (NLM). This database is viewed by medical professionals, biomedical researchers, and many other users as the authoritative source of information related to the health sciences. MEDLINE contains over 15 million references to articles from approximately 5,000 journals in 37 languages, dating back to the 1960s. In 2006, over 623,000 new citations were added to the database, and it currently grows at a rate of 2–4,000 citations daily. The subject scope of MEDLINE is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering needed by health professionals and others engaged in basic research and clinical care, public health, health policy development, or related educational activities. MEDLINE also covers life sciences vital to biomedical practitioners, researchers, and educators, including aspects of biology, environmental science, marine biology, plant and animal science as well as biophysics and chemistry.[2]

[1] http://www.ncbi.nlm.nih.gov/entrez/

[2] http://www.nlm.nih.gov/pubs/factsheets/medline.html

Each MEDLINE citation includes basic metadata information such as the title of the article, name of the authors, name of the publication, publication type, date of publication, language, etc. Of the entries added over the last decade or so, approximately 79% have English abstracts written by the authors of the articles.

PubMed is a freely-accessible Web search engine that provides access to the MEDLINE database, developed by the National Center for Biotechnology Information (NCBI) at NLM. The system provides an array of query operators that allow users to query in specific data fields (title, author, etc.) and leverage Medical Subject Headings (MeSH), drawn from NLM's controlled vocabulary thesaurus. MeSH terms are assigned manually by trained human indexers with the assistance of automated systems (Aronson et al. 2004).

A recently-revised functionality in PubMed is the "Related Links" feature. When the user examines a MEDLINE abstract, the right panel of the browser is automatically populated with titles of articles that may also be of interest, as determined by a probabilistic content similarity algorithm (Wilbur 2005)—in short, PubMed implicitly issues a query for related articles whenever a user pulls up a MEDLINE abstract to examine in detail. The goal of this feature is to unobtrusively suggest other interesting citations to facilitate knowledge discovery. The screenshot in Fig. 1 shows the arrangement of the PubMed search interface.

Article titles serve as document surrogates for presentation in the "Related Links" panel. Other space in the current PubMed interface is reserved for additional features that have yet to be deployed. Since the screen area available to this feature is fixed, we are faced with a tradeoff between quality and quantity: either show more related articles in less detail or show fewer related articles in more detail. In the simple case, this tradeoff can be implemented by controlling the amount of associated metadata that is displayed and

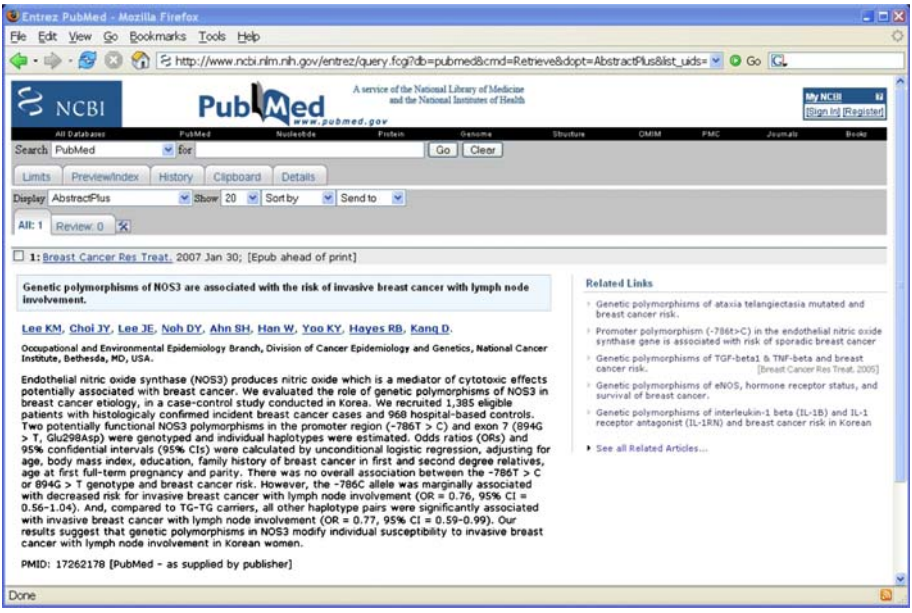

Fig. 1 Typical screenshot of the PubMed search engine as the user examines an abstract. The "Related Links" panel on the right is populated with titles of articles that may also be of interest. Other currently empty space is reserved for future functionality

truncating the title based on a fixed length quota. This work explores a more sophisticated solution based on linguistic analysis—we hypothesize that sentence compression techniques can potentially deliver the best of both worlds: by shortening article titles in a linguistically meaningful way, the interface could deliver much of the same substance in a smaller amount of space.

Ideally, the output of a compression algorithm should be both indicative and informative—properties often discussed in the context of document summarization (Afantenos et al. 2005). Indicative summaries suggest the content of an underlying information object without necessarily giving away details—often the aim is to entice users, or at least alert them to the presence of a particular information object. Movie trailers and book jackets are good examples. In contrast, an informative summary is meant to represent (and sometimes replace) the original object. Take the case of summarizing news articles: indicative summaries might mention the entities involved in a particular event, but not actually say *what happened*. An informative summary might focus on what happened, but neglect to mention the roster of participants. See (Kan et al. 2001) for an example of an attempt to introduce this distinction into summarization systems. In our application, document surrogates for related articles (displayed in the "Related Links" panel) should ideally satisfy both properties—be indicative as to capture users' interest and be informative in order to convey sufficient substance.

Ultimately, the related links feature in PubMed is designed to guide users to other articles of interest—to support information seeking or knowledge discovery. This underlying task model guides our work and provides a framework for extrinsic evaluation. The goal of the document surrogates (i.e., compressed article titles) is to facilitate interest judgments, or user decisions on whether or not to examine a particular citation. We can realistically measure the effectiveness of a compression algorithm by its ability to support such decisions.

In this context, the notion of "interest judgment" differs from the more traditional judgment of relevance that forms the basis of most retrieval applications. Ultimately, retrieval systems aim to deliver information that addresses users' information needs. However, relevance is a multi-faceted consideration that takes into account a multitude of factors—see (Mizzaro 1998) as a starting point into the rich body of literature on relevance. We do not believe it is possible to assess an article's relevance from only the title (or any short surrogate), and hence it would not be meaningful to examine relevance directly in our task context. However, an important intermediate step is the decision to examine an abstract in more detail—which we call an interest judgment. Such a decision will then cause a user to bring up more details about the article (abstract text, authors, and other metadata) in order to make a more informed decision about relevance. The related links feature in PubMed is exactly designed to elicit such interest.

Furthermore, our definition of "interest" opens the door to different types of relations beyond relevance—a citation may be interesting, not because it is potentially relevant to the present information need, but because it raises questions that the user may not have previously considered. These types of serendipitous connections underlie many significant breakthroughs in the life sciences, and PubMed aspires to assist in this process of knowledge discovery by drawing links where none previously existed.

3 Related work

Our approach to sentence compression is most similar to the work of Zajic et al. (2004) in that both use a series of linguistically-motivated trimming rules to remove inessential

fragments from the parse tree of a sentence—additional details can be found in (Dorr et al. 2003; Zajic et al. 2007, in press). This approach has proven to be highly effective at generating very short summaries of single newswire documents (i.e., the headline generation task), as evidenced by its performance at the DUC 2004 summarization evaluation. Topiary, the University of Maryland's system which integrates "parse-and-trim" techniques with topic term extraction, was among the highest scoring systems for all tasks on all measures—in some cases, even beating the performance of humans in terms of automatic metrics. Other systems that make use of similar techniques include (Mani et al. 1999; Jing 2000), and more recently, (Blair-Goldensohn et al. 2004; Conroy et al. 2005). In our approach, sentence compression is achieved by removing elements—no attempt is made to reorder material within a sentence. Since our task does not involve multiple sentences, there is no opportunity to generate output that combines fragments from different sources, for example, including one sentence as a relative clause inside another (Mani et al. 1999). Thus, we conceive of sentence compression solely as the task of selecting sentential elements (words, phrases, clauses, etc.) to remove.

Sentence compression has also been tackled with supervised machine learning techniques using a noisy-channel model. Verbose text can be viewed as the output of passing the original compressed text through a "noisy channel" that inserts additional inessential content. Given the verbose text, the system's task is to reconstruct the original message. The problem can be modeled in terms of simple word-level features, as in (Banko et al. 2000), or in terms of parse tree structures, as in (Knight and Marcu 2000; Turner and Charniak 2005). One downside of these statistical approaches is the need for annotated training data to learn model parameters. On the other hand, since trimming rules are able to exploit human linguistic insight, far less data is required for system development. Nevertheless, both methods can be viewed as complementary.

Another approach to generating very short summaries is to extract a list of topic descriptors indicative of content; examples include (Bergler et al. 2003; Zhou and Hovy 2003; Wang et al. 2005). The output of such techniques consists of, for the most part, noun phrases—as such, they are useful for telling a user what the important entities are, but less useful for conveying what actually happened. In other words, system output is indicative, but often not informative.

Although document summarization techniques have principally been applied to newswire text, there is a body of research that deals specifically with the summarization of medical documents—see (Afantenos et al. 2005) for a survey. A noteworthy example is PERSIVAL (McKeown et al. 2003; Elhadad et al. 2005), which leverages patient records to generate personalized summaries. In the genomics domain, automatic summarization techniques have also been applied to extracting GeneRIFs, concise phrases describing the function of genes (Ruch et al. 2003; Ling et al. 2006; Lu et al. 2006). In particular, Ruch et al. (2003) and Lu et al. (2006) both briefly mention methods for removing inessential elements from GeneRIFs, which are similar in spirit to our sentence compression techniques. In comparison to these cited articles, both on summarization of newswire and biomedical text, what sets our work apart is a focus on sentence compression as a tool to facilitate knowledge discovery in the context of an information retrieval system.

An important part of summarization research focuses on methodologies for evaluating system output, which can be broadly classified into two categories. In an *intrinsic* evaluation, system output is directly evaluated in terms of some set of norms—for example, fluency (Minel et al. 1997), coverage of key ideas (Paice 1990; Brandow et al. 1995), or similarity to an "ideal" summary (Kupiec et al. 1995). In particular, the last criterion has been operationalized in ROUGE (Lin and Hovy 2003), an automated metric that compares

system output to a number of human-generated "reference" summaries. The primary difficulty, however, lies in establishing an ideal reference (or a set of such texts)—summaries are generated for different purposes, and the human-centric nature of the task means that there is more than one "correct answer". Operationally, this results in low interannotator agreement on tasks such as sentence extraction (Salton et al. 1997).

In contrast to *intrinsic* evaluations, *extrinsic* evaluations attempt to measure how summarization impacts some other task. Developing realistic usage scenarios is challenging, but often the "goodness" of a summary can only be meaningfully operationalized in its "usefulness" for a particular task. One might, for example, measure how summaries impact question answering (Morris et al. 1992; Mani et al. 2002) or relevance judgments (Dorr et al. 2005). One possible hypothesis is that summaries allow users to make quicker decisions (since they have to read less), without compromising the quality of those decisions. Along these lines, our work is grounded in an information retrieval task, which allows us to assess the potential real-world impact of our sentence compression techniques.

4 Resource development

Since we are not aware of any existing resources for the sentence compression task in the biomedical domain, we devoted significant effort to creating a corpus of annotated examples. Our collection consists of article titles that have been manually shortened by domain experts. Instead of randomly sampling citations from the MEDLINE database, we leveraged the test collection developed from the TREC 2005 genomics track (Hersh et al. 2005), which fits well with our task model (finding articles of interest in the context of an ongoing search for information).

One salient feature of the TREC 2005 genomics track evaluation is its use of generic topic templates (GTTs) to capture users' information needs, instead of the typical free-text title, description, and narrative combinations used in other *ad hoc* retrieval tasks. The GTTs consist of semantic types, such as genes and diseases, that are embedded in common genomics-related information needs, as determined from interviews with real biologists. In total, five templates were developed, with 10 fully-instantiated topics for each—examples are shown in Fig. 2. Note that in some cases, the actual topics deviated slightly from the template structure (in order to accommodate real requests).

The genomics track employed a 10-year subset of the MEDLINE database (1994–2003), which totals 4.6 million citations, or approximately a third of the size of the entire MEDLINE database at the time it was collected in 2004. Each citation is identified by a unique pmid. In total, 32 groups submitted 59 runs to the task (both manual and automatic), which insured a rich, diverse pool of results. Relevance judgments were provided by an undergraduate student and a Ph.D. researcher in biology.

First, we randomly sampled 200 titles from the known list of relevant citations, and another 200 titles from the known list of irrelevant citations. These were then merged and randomized, producing a total of 400 titles, half of which were relevant according to the original assessors in the TREC task. We adopted this sampling process in order to obtain a good balance—a truly random sampling of the citation pool would yield a much larger fraction of irrelevant documents.

Next, these titles were presented to two human annotators—both were subject domain experts otherwise uninvolved with the project. Throughout this paper, we will refer to these annotators as "Lo" (Ph.D. in bioinformatics, B.S. in molecular biology) and "Ly" (Ph.D. in human genetics). They were provided both the original title and the information need

Template #1: Information describing standard [methods or protocols] for doing some sort of experiment or procedure.
methods or protocols: How to "open up" a cell through a process called "electroporation"

Template #2: Information describing the role(s) of a [gene] involved in a [disease].
gene: Interferon-beta
disease: Multiple Sclerosis

Template #3: Information describing the role of a [gene] in a specific [biological process].
gene: nucleoside diphosphate kinase (NM23)
biological process: tumor progression

Template #4: Information describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more [genes] in the [function of an organ] or in a [disease].
genes: CFTR and Sec61
function of an organ: degradation of CFTR
disease: which leads to cystic fibrosis

Template #5: Information describing one or more [mutations] of a given [gene] and its [biological impact or role].
gene with mutation: BRCA1 185delAG mutation
biological impact: role in ovarian cancer

Fig. 2 Templates and sample instantiations from the TREC 2005 genomics track

that the citation was retrieved to address (i.e., the template with concept instantiations). Both annotators were asked to generate a compressed version of the title by removing unimportant elements from the full title. After this was done, the 400 annotated pairs were divided into a development set and a held-out test set.

We attempted to align the resource development process with our task model as much as possible. To start, the TREC genomics track employed a subset of the MEDLINE database, which gave us a degree of confidence that findings could be directly applied to PubMed. Human generation of the compressed titles occurred in the context of information needs, much like the task setup in related article search (i.e., browsing the "Related Links" panel in PubMed). Furthermore, the needs are those typical of a particular user population, since they were generalized from interviews with real biomedical researchers. We believe that this collection encapsulates the original end-to-end task with great fidelity, thus enabling the results of laboratory experiments to be applicable in real-world environments.

Characteristics of our annotated data are shown in Table 1. We show the average compression ratio and the average length reduction (with standard deviation) both in terms of characters and words. Compression ratio is computed as the length of the compressed sentence divided by the length of the original sentence, averaged across the entire data set of 200 sentences. Thus, the smaller the number, the shorter the output is. We note that there appear to be more opportunities for compression in the development set (given the lower compression ratios). Abstract titles in the development set averaged 102.3 characters, or 13.34 words; abstract titles in the test set averaged 104.0 characters, or 13.60 words.

5 A syntactic approach to compression

We adopt a syntactic approach to sentence compression through the use of linguistically-motivated trimming rules that remove fragments of parse trees. Our work employs the

Table 1 Characteristics of the human-generated compressed abstract titles

Annotator	Characters		Words	
	Ratio	Reduction	Ratio	Reduction
Development set				
Lo	0.564	46.4 ± 28.4	0.543	6.34 ± 3.81
Ly	0.575	48.0 ± 34.9	0.567	6.40 ± 4.61
Test set				
Lo	0.601	43.6 ± 26.5	0.587	5.99 ± 3.77
Ly	0.652	39.0 ± 30.6	0.644	5.25 ± 4.12

The table shows compression ratio and average length reduction (with standard deviation) at the character and word levels: development set on top and test set on bottom. Length of complete abstract titles: 102.3 characters, 13.34 words (for development set) and 104.0 characters, 13.60 words (for test set)

Stanford Parser (Klein and Manning 2003).[3] Although it was not originally designed to parse text in the biomedical domain, experimental results show that a syntactic approach is nevertheless effective for compressing abstract titles in the biomedical domain (more on this in Sect. "Discussion"). The techniques described here are not tied to a particular parser and will function with any system that utilizes the Penn Treebank conventions.

From the development set, we came up with seven linguistically-motivated rules—many of these are similar to those discussed in (Zajic et al. 2004). These rules are described below, arranged roughly in increasing order of complexity. An example of each is shown in Table 2; those examples illustrate the application of each rule in isolation.

- *Subtitles:* Subtitles, denoted with a colon or consecutive dashes, are removed. This is accomplished with simple regular expression patterns prior to parsing. Examination of titles from the development set suggests that the recognition of colons and consecutive dashes is sufficient to identify subtitles.
- *Determiners (DT):* Determiners are removed. All terminals in the parse tree assigned the part-of-speech tag DT are considered determiners.
- *Participial and Gerund phrases (VBG):* Participial and gerund phrases are removed. These phrases are recognized by traversing the parse tree and identifying VPs headed by tokens tagged as VBG, i.e., the structure $[_{\mathrm{VP}}[_{\mathrm{VBG}}\,\mathrm{X}]\cdots]$. Resulting empty nodes (e.g., dangling prepositions) are also removed.
- *Serial PPs:* In a sequence of prepositional phrases (sharing a common ancestor), all but the first are removed. Prepositional phrases are recognized by traversing the parse tree and identifying constituents with the structure $[_{\mathrm{PP}}[_{\mathrm{IN}}\,\mathrm{X}]\cdots]$. See Sect. "Discussion" for a detailed discussion regarding prepositional phrase attachment issues with the Stanford Parser.
- *Nested PPs:* Any prepositional phrase that is embedded three or more levels deep is removed. As with the previous rule, PPs are identified by traversing the parse tree and identifying structures of the following form: $[_{\mathrm{PP}}[_{\mathrm{IN}}\,\mathrm{X}]\cdots]$ See also Sect. "Discussion" for issues related to prepositional phrase attachment.
- *Conjoined NPs:* For conjoined noun phrases, the second conjunct is removed. This rule is motivated by Zajic et al. (2007, in press), who observed that the first conjunct in a

[3] Version 1.5.1, downloaded from http://www-nlp.stanford.edu/software/; default settings were used in all experiments.

Table 2 Example of all trimming rules. All rules are applied in isolation

Rule	Example
Subtitle	Structure and dynamics of the GABA binding pocket~~: A narrowing cleft that constricts during activation.~~ (PMID: 11150321)
DT	~~The~~ effectiveness of ~~a~~ peripatetic allergy nurse practitioner service in managing atopic allergy in general practice — ~~a~~ pilot study. (PMID: 10779298)
VBG	Mutation analysis in a small cohort of New Zealand patients [$_{\text{VBG}}$ ~~originating from the United Kingdom~~] demonstrates genetic heterogeneity in familial hypercholesterolemia. (PMID: 11040093)
Serial PP	Electroporation-mediated interleukin-12 gene therapy [$_{\text{PP}}$ for hepatocellular carcinoma] [$_{\text{PP}}$ ~~in the mice model~~]. (PMID: 11221826)
Nested PP	Semiquantitative immunoblot analysis [$_{\text{PP}}$ of nm23-H1 and -H2 isoforms [$_{\text{PP}}$ ~~in adenocarcinomas of the lung~~]]: prognostic significance. (PMID: 10792783)
Conjoined NP	[$_{\text{NP}}$ [$_{\text{NP}}$ Extraction] ~~and~~$_{\text{CC}}$ [$_{\text{NP}}$ ~~isolation~~] of linear alkylbenzenesulfonate] ~~and~~$_{\text{CC}}$ [$_{\text{NP}}$ ~~its sulfophenylcarboxylic acid metabolites~~] from fish samples. (PMID: 10575968)
Simple NP	Acetylation of [$_{\text{NP}}$ a ~~specific~~$_{\text{JJ}}$ promoter$_{\text{NP}}$ nucleosome$_{\text{NP}}$] accompanies activation of [$_{\text{NP}}$ the epsilon-globin$_{\text{JJ}}$ gene$_{\text{NP}}$]. (PMID: 11158302)

$[_{\text{NP}}[_{\text{NP}} \cdots]$ and$_{\text{CC}}[_{\text{NP}} \cdots]]$ structure is often more important. Noun phrases containing conjoined adjectives are left untouched.

- *Simple NPs:* Adjectives in NPs are removed, unless their heads are "lightweight". For example, contrast "a specific$_{\text{JJ}}$ promoter$_{\text{NP}}$ nucleosome$_{\text{NP}}$" and "the epsilon-globin$_{\text{JJ}}$ gene$_{\text{NP}}$": whereas the adjective can be removed without significantly affecting content in the first case, the token tagged as adjective ("epsilon-globin") actually conveys most of the content in the second NP. We term the head (rightmost) noun lightweight in the second case and approximate its "weight" using inverse document frequency (*idf*), a commonly-employed measure in information retrieval. Thus, a threshold associated with this rule specifies the weight of a head noun above which modifying adjectives are removed.

The above trimming rules were formulated after examining sentence pairs in the development set and recognizing opportunities to remove inessential fragments of titles based on syntactic structure. This was not accomplished in a systematic way, and no doubt there are other opportunities that can be exploited. Nevertheless, as our evaluations show, these rules provide the basis for an effective sentence compression algorithm. The order in which the trimming rules are sequentially applied in our final implementation was guided by the evaluation of individual rules, described in the next section.

6 Automatic evaluation

Although end-to-end task-based evaluations provide the best method for assessing information systems, the large amount of manual effort typically required for such evaluations precludes

using them for system development. In human language technologies, researchers often employ a paradigm based on automatic metrics for system development, capped with a summative evaluation. Our work follows in this model. This section describes a series of experiments that characterize the effectiveness of our syntactic trimming techniques using automated methods. A manual evaluation is detailed in Sect. "Manual Evaluation".

6.1 Evaluation methodology

Automatic evaluation methods are attractive because they enable quick experimental turnaround, thereby facilitating rapid exploration of the solution space. An easily quantifiable performance metric provides researchers with an objective function over which they can optimize. Once an automatic metric has been validated—that is, demonstrated to correlate with human preferences—the measure can be exploited for system development.

In the language processing community, researchers have developed a family of automatic metrics based on the idea of comparing system output to one or more human-generated references. Similarity to these "gold standards", according to different content overlap metrics, can quantify the quality of system output—this represents a well-established evaluation methodology in the language processing community. Two such commonly-used metrics are BLEU (Papineni et al. 2002) for machine translation and ROUGE (Lin and Hovy 2003) for document summarization. Both rely on computing *n*-gram overlap between system output and human references (manually-translated sentences and human-generated summaries, respectively), but the details differ. In general, BLEU is a precision-oriented metric that places heavy emphasis on fluency, i.e., checking to make sure that the system output is "good English". This is important in evaluating machine translations since automated systems have a tendency to produced garbled sentences. ROUGE, a metric developed for document summarization tasks, on the other hand, focuses on recall of content, i.e., the presence or absence of certain topic terms. Because most modern document summarization systems are extractive, generation of disfluent output is not as severe a problem.[4] Thus, summarization metrics emphasize the inclusion of key facts or concepts in the system output, as measured in terms of *n*-gram overlap.

Given these considerations, we decided that BLEU is the more appropriate metric for our sentence compression task—primarily because we are concerned with the fluency of system output. Conceptually, sentence compression can be viewed as "translating" from verbose English into succinct English. According to previous work (Lin 2004), ROUGE-1 recall correlates best with human judgments on the headline generation task—constructing a very short summary (less than 75 characters) from a single newspaper article. This is the closest summarization analog to our task, but we believe that ROUGE-1 is not an appropriate metric. The measure focuses on unigram overlap, i.e., the presence of content words in system output, and is not sensitive to fluency considerations; a grammatical sentence and a random sequence of the same words would receive identical scores. This characteristic does not fit with our desire to assess the grammatical correctness of system output, which is an important concern since removing certain portions of the parse tree may yield ungrammatical output. BLEU, on the other hand, is better able to model fluent English text since it takes into account *n*-grams of different lengths. Although the metric considers only the surface properties of machine output and lacks even a rudimentary model of syntax and semantics, it has proven highly effective in

[4] On a sentence-by-sentence basis, that is. Discourse-level properties such as coherence of text are important issues, but much harder to evaluate and quantify.

Table 3 BLEU scores of one annotator's output using the other as a reference, on both the development set and the test set

Output	Reference	Dev	Test
Lo	Ly	0.510	0.550
Ly	Lo	0.513	0.544

These values characterize the agreement between the two human annotators

guiding research in machine translation. Despite its deficiencies, BLEU has enabled rapid progress in translation technology over the past few years.

To get a sense of how much the assessors agreed with each other, we computed BLEU scores, using one as the "system output" and the other as the reference. These results are shown in Table 3. Note that these values are lower than many of the BLEU scores reported for our compression algorithm because only one set of references is used; in all other experiments, system output is evaluated against both sets of human references.

6.2 Application of individual rules

We first examined each rule in isolation. Table 4 shows results for all rules expect for the simple NP rule, which requires an additional parameter and is therefore discussed separately. For each rule, we note the number of titles that triggered the rule (out of 200 in each

Table 4 Compression effectiveness of each rule in isolation with the exception of the simple NP rule

Rule	#	Characters		Words		BLEU
		Ratio	Reduction	Ratio	Reduction	
Development set						
Untrimmed	–	1.000	0.00	1.000	0.00	0.538
Subtitle	39	0.525	49.92 ± 27.36	0.518	6.54 ± 3.61	0.579
DT	103	0.955	4.64 ± 2.57	0.899	1.40 ± 0.66	0.555
VBG	20	0.584	47.20 ± 23.56	0.573	6.40 ± 3.10	0.541
Serial PP	40	0.587	39.85 ± 21.78	0.564	5.70 ± 3.12	0.527
Nested PP	41	0.748	32.85 ± 23.07	0.723	4.76 ± 2.99	0.540
Conjoined NP	73	0.722	29.30 ± 17.89	0.703	4.10 ± 2.53	0.514
Test set						
Untrimmed	–	1.000	0.00	1.000	0.00	0.593
Subtitle	38	0.510	53.45 ± 29.83	0.529	6.82 ± 4.20	0.619
DT	109	0.956	4.61 ± 2.68	0.907	1.34 ± 0.74	0.600
VBG	26	0.656	36.35 ± 29.24	0.683	4.31 ± 3.98	0.586
Serial PP	42	0.688	34.86 ± 28.18	0.670	4.93 ± 3.93	0.596
Nested PP	36	0.778	28.50 ± 20.16	0.759	4.22 ± 3.05	0.591
Conjoined NP	54	0.763	25.93 ± 16.22	0.753	3.57 ± 2.44	0.575

The second column shows number of abstract titles (out of 200) that triggered the rule. The last column shows the BLEU score of compressed output. Middle columns show average compression ratio and average length reduction (with standard deviation), in terms of characters and words. Data from development set shown on top, and data from test set shown on bottom. Length of complete abstract titles: 102.3 characters, 13.34 words (for development set) and 104.0 characters, 13.60 words (for test set)

set) in the second column of the table. This value quantifies the prevalence of each phenomena. The next four columns show average compression ratio and average length reduction (with standard deviation), both at the character and word levels. These values are computed over affected titles only (the set of abstract titles that triggered the rule). The BLEU scores were computed across the entire data sets (all 200 sentences), using both "Lo" and "Ly" as the references.

How is one supposed to interpret these results? The effectiveness of each rule is quantified in two ways: the amount of compression achieved and how "good" the resulting output is (compared to human-generated references using the BLEU metric). Thus, each rule represents a tradeoff along these two dimensions. Leaving the abstract titles untrimmed (the first row in Table 4) represents a baseline. Not surprisingly, application of the trimming rules in most cases raises the BLEU score, indicating that the results are closer to the references than the original full title in terms of *n*-gram content. In most cases, gains observed in the development set carried over to the held-out test set, although the magnitude of the improvements were smaller (but recall that the human-annotated gold standards suggest fewer opportunities for trimming in the test data). In general, the scope of the rules (i.e., number of affected titles) and the degree of compression were comparable.

The performance of the simple NP trimming rule in isolation is shown in Fig. 3. Corpus statistics required for the *idf* calculation were extracted from the 10-year MEDLINE collection used in the TREC 2005 genomics track. We varied the *idf* threshold from 2.0 to 10.0 in increments of 0.5, and obtained a plot that relates the BLEU score to the average compression ratio at the character level. For this graph, average compression ratio is computed on all the titles, even those that were unaffected by the rule (since the threshold controls how many titles trigger the rule). The corresponding graph for average compression ratio at the word level looks nearly identical, and is not shown here. It is interesting to note that no threshold actually increases the BLEU score above the baseline

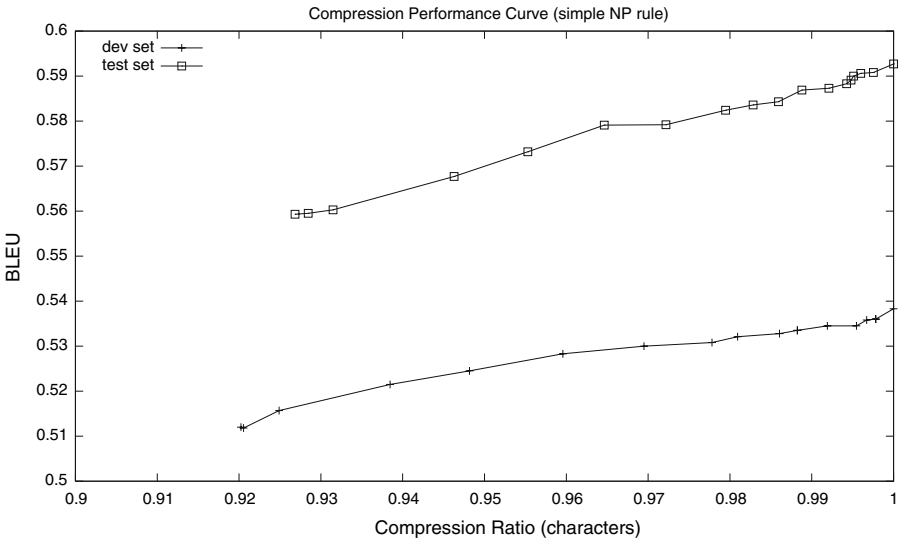

Fig. 3 Performance curve of the simple NP rule with *idf* threshold ranging from 2.0 to 10.0 in 0.5 increments

(no compression), indicating that users engage in more complex behavior than simply removing modifiers of noun phrases based on *idf* values. Trends observed in the development set carry over to the held-out data, although the rule yields less compression.

Given that each rule represents a tradeoff between output quality and compression ratio, how can one assemble a complete compression algorithm for MEDLINE abstract titles? In response, we examine sequential application of the trimming rules, which yields curves that characterize the tradeoffs mentioned. In Sect. "Manual evaluation", we report results on manual evaluation of system output at two specific points on this tradeoff curve.

6.3 Sequential application of multiple rules

After examining each rule in isolation, we experimented with applying the rules sequentially. The rules were ordered based on BLEU scores on the development set (Table 4): subtitle, DT, VBG, nested PP, serial PP, conjoined NP, and finally simple NP. Results are displayed in Fig. 4; the top plot shows compression ratios computed at the character level, and the bottom plot shows compression ratios computed at the word level. The horizontal dotted lines denote the BLEU scores of the original (untrimmed) article titles. The left tails of the curves represent the application of the simple NP rule with different *idf* thresholds (we varied the parameter from 2.0 to 10.0 in increments of 0.5). The labels indicate the points at which each rule is applied (except for the simple NP rule to reduce clutter). In these plots, points closer to the upper left hand corner are "better", in the sense that we desire large reductions in length while maintaining fidelity to the human generated references.

We notice that successive application of individual rules is subjected to a "diminishing returns" effect. The amount of compression achieved with the application of multiple rules is strictly less than the sum of the compression achieved by individual rules. This occurs because earlier rules can eliminate opportunities for later rules to apply. Take the example of prepositional phrases or conjoined NPs inside a gerund phrase. Since the VBG rule applies earlier, the entire phrase would have already been eliminated.

The same general characteristics are observed in both the development and test sets: BLEU scores initially rise and then drop as the rules more and more aggressively trim away parts of the structure. Untrimmed sentences (right edges of the graphs) serve as a baseline—but note that it is possible for our compression algorithm to perform worse if too much is removed from the abstract titles. The dip in the performance curve on the test set can be attributed to the relative performance of the PP trimming rules: one was found to be more effective in the development set, but the reverse turned out to be true in the test set.

How can one interpret these results, especially since BLEU scores do not correspond to any quantity that humans have an intuition for? There are two responses to this question: first, although it is difficult to map absolute scores to a particular level of performance, relative differences in BLEU are meaningful—in that they tell us if one variant is "better" than another (in terms of matching human references).

However, the more appropriate response is to acknowledge the limitations of automatic scoring metrics. Ultimately, our goal is to develop information systems that are useful for humans, and one way of operationalizing "usefulness" is in terms of task performance. Therefore, we believe that the question "How good is a BLEU score of 0.452?" is not pertinent. Rather, we must ask if particular techniques can better assist humans in accomplishing real-world tasks. System development, as guided by automatic metrics, only

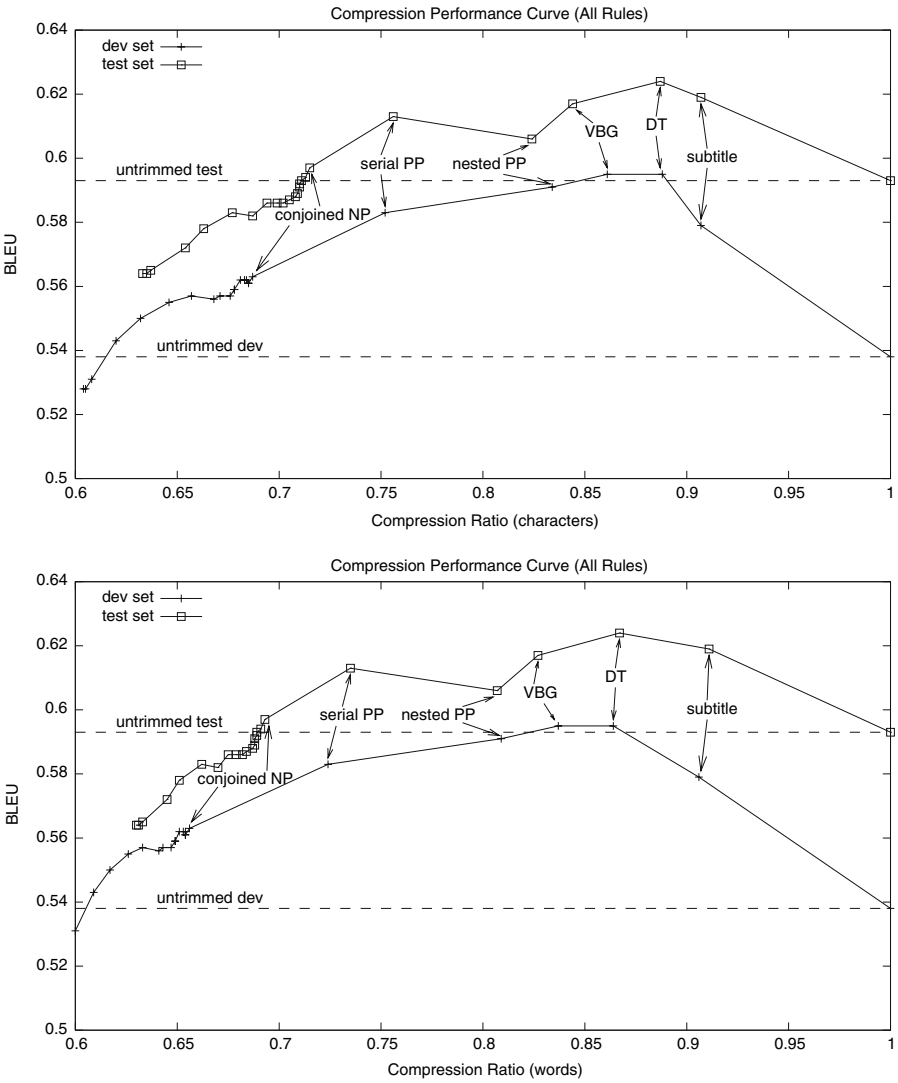

Fig. 4 Plots of average compression ratio versus BLEU score, with sequential application of all rules: subtitle, DT, VBG, nested PP, serial PP, conjoined NP, and simple NP (*idf* threshold ranging from 2.0 to 10.0 in 0.5 increments). Average compression ratio in terms of characters shown on top, in terms of words shown on bottom. Horizontal dotted lines show BLEU scores of untrimmed abstract titles

serves as a stepping stone to extrinsic task-based evaluations. In the next section, we report results from exactly such a study.

7 Manual evaluation

In our initial experiments, BLEU primarily served as a formative tool to guide system development. We then conducted a summative task-based evaluation to assess the

Table 5 Fluency and content judgments by three assessors, on a scale of 1–5 (1 = worst, 5 = best); means and standard deviations are reported

Output	Ratio	Fluency			Content		
		Lo	Ly	Wa	Lo	Ly	Wa
Lo	0.564	3.9 ± 0.95	4.0 ± 1.06	4.0 ± 0.98	3.5 ± 0.96	4.2 ± 0.76	3.6 ± 0.95
Ly	0.575	4.6 ± 0.76	3.0 ± 1.57	4.5 ± 0.87	3.2 ± 1.36	4.3 ± 0.69	3.7 ± 1.06
Variant A	0.687	4.5 ± 0.82	3.4 ± 1.66	4.0 ± 0.93	3.4 ± 1.61	4.0 ± 0.98	3.8 ± 1.18
Variant B	0.563	3.4 ± 1.50	2.8 ± 1.45	3.1 ± 1.36	2.4 ± 1.66	3.5 ± 1.26	3.2 ± 1.19

The second column indicates the mean character compression ratio for each condition

usefulness of our sentence compression algorithm. At the same time, we also collected human judgments about the intrinsic quality of the compressed output.

Recall from Sect. "Motivation" that we situate sentence compression in the context of related article search in PubMed—in particular, as a method for efficient presentation of items that may be of interest to the user. In this context, the end goal of our sentence compression algorithm is to support interest judgments, that is, a user's decision to examine a citation in detail. The default condition is to show the full title, which serves as a baseline. If the output of our sentence compression algorithm is able to provide the same level of decision support (in terms of accuracy of interest judgments), but with a smaller amount of text, then we can claim to have improved on the baseline. The ability to convey much of the same content in fewer words can be leveraged in two different ways: PubMed can use a smaller screen area for "Related Links", thereby freeing up space or other content elements, or PubMed can display more related articles in the same amount of on-screen space.

Three subject domain experts uninvolved with system development were recruited as assessors. Two of them ("Lo" and "Ly") were the same individuals involved in creating our training and test sets. The third individual ("Wa") was not involved in any other aspect of the project.

Our experiments involved 100 article titles randomly sampled from the development set.[5] For each title, we randomly assigned one of four conditions to the trimmed output (25 examples each), described below. The average compression ratio (at the character level) of each condition is shown in the second column of Table 5.

- *Lo:* Manually compressed titles by the annotator "Lo".
- *Ly:* Manually compressed titles by the annotator "Ly".
- *Variant A*: Application of the following rules to the original titles: subtitle, DT, VBG, nested PP, serial PP, and conjoined NP. Note that these titles were longer than the human gold standards on average.
- *Variant B:* Variant A plus the application of the simple NP rule, with an *idf* threshold of 2.0 (aggressive compression, same as the left tail on the plots in Fig. 4). On average, these titles were approximately the same length as the human references.

[5] We avoided using the held-out test data so that we can continue developing the trimming algorithm in the future. This does not impact our findings, since the development-test division is meaningless from our assessors' point of view.

The assessment proceeded in two rounds. In the first round, assessors were provided with the information need and the trimmed title of the abstract. Naturally, they did not know the source of the trimmed titles (i.e., which condition). Assessors were asked to rate the fluency of the title on a scale of 1–5 (1 = worst, 5 = best). They were also asked if they would read the citation in response to the information need (i.e., a judgment of interest).

In the second round, assessors were provided with the information need, the original title, the trimmed title, and their interest judgment from the first round (i.e., response to "Would read abstract?"). They were asked to rate the content of the trimmed title, in terms of capturing essential elements from the full title, on a scale of 1–5 (1 = worst, 5 = best). In addition, they were asked if they would now read the citation given the full title.

This two-phase setup was designed to evaluate both the intrinsic quality of the compressed titles and their effectiveness in a task context. Fluency and content judgments characterize the inherent quality of the compressed titles, while the interest judgments ground our evaluation in a real-world scenario. We were especially concerned about differences in interest judgments from round one and round two. If both responses were the same, we can conclude that the article title was shortened successfully, i.e., the process did not interfere with task performance. A "yes to no" flip provides evidence that the compression process created a false impression of interest. A "no to yes" flip, on the other hand, provides evidence that essential elements from the title were mistakenly removed.[6]

Fluency and content judgments from our human assessors are provided in Table 5, which shows ratings given by all three assessors on all four conditions (mean and standard deviation). Results suggest that variant A is not any more disfluent than the human-compressed gold standards: two of the three assessors actually placed the machine-generated output ahead of one of the gold standards. Note, however, that variant A titles were on average longer than human references. Machine-generated output of comparable compression (variant B) was found to be consistently less fluent than the other conditions. Similar trends are observed for content ratings: variant A appears to be as good as human output, whereas variant B is less so. Overall, there appears to be much variability in these judgments, suggesting that differences exist in the assessors' interpretation of the task. In particular, we note that Lo preferred Ly's output to Lo's own, in terms of both fluency and content. Similarly, Ly preferred Lo's output to Ly's own (again, both fluency and content). We currently have no reasonable explanation for this observation.

How does the quality of trimmed titles affect users' task performance? The answer can be found in the tally of flips in interest judgments, as shown in Table 6. These results appear to suggest that compressed titles have relatively minimal impact on users' ability to decide whether they want to examine a citation. There does not appear to be much difference between variant A and either one of the human-generated compressions, although variant B titles caused more flips.

Another way of organizing the results is to compare users' interest judgments on the full title (from the second round) with their judgments on compressed titles (from the first round). This is similar to the *consistency test* employed in the TIPSTER SUMMAC evaluations (Mani et al. 2002). Results can be analyzed in terms of the contingency table shown in Table 7, from which we can compute standard aggregate statistics:

[6] Note that since the evaluated set was constructed with a balanced number of relevant and irrelevant articles (as deterimined by the original genomics track assessors), the consistency measures are not distorted by class imbalance issues.

Table 6 Flips in interest judgments (out of a maximum of 25 for each condition) by three assessors based on compressed (round 1) and full titles (round 2)

Output	Lo		Ly		Wa	
	Y → N	N → Y	Y → N	N → Y	Y → N	N → Y
Lo	2	0	1	1	2	0
Ly	0	1	1	0	0	0
Variant A	0	3	1	0	1	1
Variant B	1	2	2	3	2	2

Table 7 Contingency table for interest judgments on full titles and on compressed variants

Full	Compressed	
	Interesting	¬Interesting
Interesting	True positive (TP)	False negative (FN)
¬Interesting	False positive (FP)	True negative (TN)

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$
$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$
$$F\text{-score} = 2 \times \text{Precision} \times \text{Recall}/(\text{Precision} + \text{Recall})$$

Results of this analysis are presented in Table 8. According to the three assessors, variant A titles perform on par with the human-compressed versions, although humans are able to achieve more compression. Variant B titles, which are approximately the same length as the human-generated references, appear to perform worse in terms of precision, recall, and *F*-score. Note that the absolute performance achieved by human-compressed titles and the output of variant A is very high—almost perfect in many cases. This confirms our basic premise that inessential fragments from MEDLINE article titles can be removed without affecting the substance of what is conveyed.

What are the implications of our findings? Our syntactic compression algorithm is able to shorten abstract titles by approximately 30% without noticeably affecting task performance (variant A). This translates into less material for the user to read, or alternatively, over 40% more content per unit area. In the same space that it takes to display five full

Table 8 P(recision), R(ecall), and *F*(-score) in recognizing interesting citations with four different compressed output, according to three different assessors

Output	Lo			Ly			Wa		
	P	R	*F*	P	R	*F*	P	R	*F*
Lo	0.80	1.00	0.89	0.96	0.96	0.96	0.91	1.00	0.95
Ly	1.00	0.88	0.93	0.93	1.00	0.97	1.00	1.00	1.00
Variant A	1.00	0.73	0.84	0.94	1.00	0.97	0.94	0.94	0.94
Variant B	0.50	0.33	0.40	0.86	0.80	0.83	0.82	0.82	0.82

abstract titles, we can now display seven. We believe this is a significant result because it provides users access to more potentially interesting articles without requiring them to read more text.

8 Discussion

At a broader level, we believe that this work is significant in two ways. First, the application of syntactic compression techniques in the biomedical domain raises interesting questions about the domain portability of existing language processing tools. Second, we view this work as a case study highlighting the importance of task-based evaluations in grounding summarization technology. This section elaborates on both points.

8.1 Domain adaptation (Or lack thereof)

Issues surrounding the portability of text processing algorithms have recently gained interest in the research community. Due to the availability of corpora and other resources, most modern statistical tools are trained on newswire text, and hence specialized for processing text from that genre, even though many other types of text are worth exploring. Thus, an important consideration in the development of language technologies is its ability to generalize across different domains and genres of text. In this work, we demonstrate that sentence compression techniques originally developed for news articles can be effectively applied to compress MEDLINE article titles. In some ways, this result is somewhat surprising, as we explain below.

First, the biomedical domain offers significant challenges to off-the-shelf parsers. The lexical overlap between MEDLINE abstracts and typical news corpora is surprisingly small (Smith et al. 2005), which creates challenges for parsers—see, for example, (Clegg and Shepherd 2005; Grover et al. 2005; Lease and Charniak 2005). As a specific example, prepositional phrase attachment is problematic for statistical parsers trained on newswire text, since noun phrase heads are often unknown lexical items in the biomedical domain. Consider examples taken from Table 2:

(1) Electroporation-mediated interleukin-12 gene therapy [$_{PP}$ for hepatocellular carcinoma][$_{PP}$ in the mice model].
(2) Semiquantitative immunoblot analysis [$_{PP}$ of nm23-H1 and -H2 isoforms [$_{PP}$ in adenocarcinomas of the lung]]: prognostic significance.

These abstract titles are typical of those in MEDLINE—characterized by sequences of consecutive prepositional phrases. The examples above are annotated with actual structures assigned by the Stanford Parser, which are correct in both cases. The attachment of PPs, whether to the immediately preceding noun head—the case with example (2)—or another head higher up in the structure—the case with example (1)—is a complex decision that often requires semantic knowledge. Often, the Stanford Parser is incorrect in its choice of PP attachment point, as illustrated by the following parse:

(3) Induction [$_{PP}$ of cell cycle arrest and morphological differentiation [$_{PP}$ by Nurr1 and retinoids]][$_{PP}$ in dopamine MN9D cells].

Rather, the correct structure should be:

(4) Induction [$_{PP}$ of cell cycle arrest and morphological differentiation][$_{PP}$ by Nurr1 and retinoids][$_{PP}$ in dopamine MN9D cells].

Nevertheless, our nested and serial PP rules appear to be insensitive to these errors because they were engineered by examining Stanford Parser output. Since the parser appears to make systematic errors, we are still able to capture generalizations—albeit these rules may not represent linguistically valid generalizations in the biomedical domain. In example (3), the serial PP rule removes the PP "in dopamine MN9D cells", which appears to yield a reasonable compression.

Second, MEDLINE article titles are quite different from sentences that occur in newswire text—most of the time, titles are not even complete sentences. Since titles are often noun phrases or verb phrases, our approach must not only cope with out-of-domain effects (most notably, unknown lexical items), but also stylistic differences. Experiments suggest that the syntactic trimming approach is also capable of handling such divergences, given a set of rules developed specifically for the biomedical domain. It is noteworthy that respectable performance is achieved in our application without any domain adaptation.

8.2 Grounding summarization in real-world tasks

This work serves as a case study illustrating the importance of grounding summarization tasks in real-world user scenarios. To a human, a fluency score of four (out of five) is not particularly meaningful, and neither is a 0.314 BLEU score. However, quantifying performance in terms of decision-making accuracy on compressed titles (as compared to the full titles) is informative because it illustrates how summarization techniques assist the user's end task, that of knowledge exploration and information gathering. In general, we believe that information presentation issues provide a general framework for task-based evaluation of summarization systems; see also, (Mani et al. 2002; Dorr et al. 2005) for similar setups.

We would like to end this section with a discussion of our underlying task model. In most retrieval tasks, the assumption is that the user issues a query to a search engine and obtains a ranked list of documents that are potentially relevant. This output then serves as the starting point to browsing, selection, examination, and query reformulation mechanisms that may ultimately lead to the satisfaction of the information need. However, this traditional query-centered model does not describe the only possible pattern of user-system interactions. In particular, PubMed attempts to draw connections between articles in MEDLINE by unobtrusively displaying titles that may be of interest. This mechanism provides users with another device for exploring the information space. In fact, previous simulation studies have shown that such a feature can improve performance, as measured by traditional ranked retrieval metrics (Wilbur and Coffee 1994; Smucker and Allan 2006). Although we focus primarily on the related links features in PubMed, our syntactic compression techniques are equally applicable to other components in the retrieval environment, e.g., summarizing ranked lists so that more results can be displayed on any given Web page.

9 Future work

With respect to syntactic compression in the biomedical domain, there are two distinct threads of future work worth exploring—improvements to the compression algorithm and application of similar techniques to related problems. We briefly discuss each.

The simple NP rule is currently the only rule that is parameterized—in this case, an *idf* threshold. The same idea could be applied to other rules.[7] For example, the system could use a threshold to determine if the head of a prepositional phrase was "lightweight", and then factor in this evidence to determine if the PP was removable. In the same way, the conjoined NP rule could also be subjected to this modification.

However, the introduction of parameterized rules adds additional complexity to rule ordering. Currently, trimming rules are applied sequentially in a fixed order (based on individual performance in isolation). This neglects possible interaction effects between rules, which would certainly increase with the introduction of parameters. Zajic (2007) explored a solution to a similar problem by allowing multiple simultaneous rule applications, and then developing a mechanism to select among the multiple compressed candidates. We believe that the same idea can be applied here.

Beyond applications in information retrieval, sentence compression techniques can also be used for other tasks in the biomedical domain. The extraction of GeneRIFs is one such possibility.[8] GeneRIFs are concise phrases describing a function of a gene, explicitly linked to the Entrez Gene database. The extraction and linking of these descriptions is important for biologists and other researchers, since such information would otherwise be scattered in many disparate articles and sources. GeneRIFs by design are limited to 255 characters,[9] so brevity is highly desired. As previously discussed, summarization techniques have been successfully applied to extracting GeneRIFs (Ling et al. 2006; Lu et al. 2006)—these systems could additionally benefit from the syntactic compression techniques discussed in this article, to eliminate inessential material from the extracted phrases.

10 Conclusion

Like much previous work, this paper examines the task of sentence compression. However, our perspective is novel in two different ways: we explore the problem in the domain of biomedicine and within the context of an information retrieval task. The contributions of this study are two-fold: first, we demonstrate that the syntactic trimming approach, which has proven effective in the newswire domain, is portable to the biomedical domain. Second, our work highlights the importance of extrinsic evaluations and grounding summarization in real-world tasks. It is our hope that we can eventually transition summarization technology into the PubMed search engine.

Acknowledgments We would like to thank our annotators and assessors for their efforts, without which this study would not be possible. Our thanks also go out to the anonymous reviewers, whose comments and criticisms have made this article significantly better. For this work, JL was funded in part by the National Library of Medicine, where he was a visiting research scientist during the summer of 2006. WJW is supported by the Intramural Research Program of the NIH, National Library of Medicine. JL would also like to thank Esther and Kiri for their kind support.

[7] We would like to thank an anonymous reviewer for pointing this out.

[8] We would again like to thank an anonymous reviewer for pointing this out.

[9] http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html

References

Afantenos, Stergos, Karkaletsis, Vangelis, & Stamatopoulos, Panagiotis (2005). Summarization from medical documents: A survey. *Artificial Intelligence in Medicine, 33*(2), 157–177.

Aronson, Alan R., Mork, James G., Gay, Clifford W., Humphrey, Susanne M., & Rogers, Willie J. (2004). The NLM indexing initiative's medical text indexer. In *Proceedings of the 11th World Congress on Medical Informatics (MEDINFO 2004)* (pp. 268–272). San Francisco, California.

Banko, Michele, Mittal, Vibhu, & Witbrock, Michael (2000). Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)* (pp. 318–325). Hong Kong, China.

Bergler, Sabine, Witte, René, Khalife, Michelle, Li, Zhuoyan, & Rudzicz, Frank (2003). Using knowledge-poor coreference resolution for text summarization. In *Proceedings of the HLT/NAACL 2003 Text Summarization Workshop and Document Understanding Conference (DUC 2003)* (pp. 85–92). Edmonton, Alberta.

Blair-Goldensohn, Sasha, Evans, David, Hatzivassiloglou, Vasileios, McKeown, Kathleen, Nenkova, Ani, Passonneau, Rebecca, Schiffman, Barry, Schlaikjer, Andrew, Siddharthan, Advaith, & Siegelman, Sergey (2004). Columbia University at DUC 2004. In *Proceedings of the 2004 Document Understanding Conference (DUC 2004) at HLT/NAACL 2004* (pp. 23–30). Boston, Massachusetts.

Brandow, Ronald, Mitze, Karl, & Rau, Lisa F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management, 31*(5), 675–685.

Clegg, Andrew B., & Shepherd, Adrian (2005). Evaluating and integrating Treebank parsers on a biomedical corpus. In *Proceedings of the ACL 2005 Workshop on Software*. Ann Arbor, Michigan.

Conroy, John M., Schlesinger, Judith D., & Stewart, Jade Goldstein (2005). CLASSY query-based multi-document summarization. In *Proceedings of the 2005 Document Understanding Conference (DUC-2005) at NLT/EMNLP 2005*. Vancouver, Canada.

Dorr, Bonnie J., Zajic, David, & Schwartz, Richard (2003). Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT/NAACL 2003 Text Summarization Workshop and Document Understanding Conference (DUC 2003)* (pp. 1–8). Edmonton, Alberta.

Dorr, Bonnie J., Monz, Christof, President, Stacy, Schwartz, Richard, & Zajic, David (2005). A methodology for extrinsic evaluation of text summarization: Does ROUGE correlate? In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Ann Arbor, Michigan.

Elhadad, Noemie, Kan, Min-Yen, Klavans, Judith, & McKeown, Kathleen (2005). Customization in a unified framework for summarizing medical literature. *Journal of Artificial Intelligence in Medicine, 33*(2), 179–198.

Grover, Claire, Lascarides, Alex, & Lapta, Mirella (2005). A comparison of parsing technologies for the biomedical domain. *Natural Language Engineering, 11*(1), 27–65.

Hersh, William, Cohen, Aaron, Yang, Jianji, Bhupatiraju, Ravi, Roberts, Phoebe, & Hearst, Marti (2005). TREC 2005 genomics track overview. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*. Gaithersburg, Maryland.

Jing, Hongyan (2000). Sentence reduction for automatic text summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'00)*. Seattle, Washington.

Kan, Min-Yen, McKeown, Kathleen R., & Klavans, Judith L. (2001). Domain-specific informative and indicative summarization for information retrieval. In *Proceedings of the 2001 Document Understanding Conference (DUC 2001) at SIGIR 2001*. New Orleans, Louisiana.

Klein, Dan, & Manning, Christopher D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)* (pp. 423–430). Sapporo, Japan.

Knight, Kevin, & Marcu, Daniel (2000). Statistics-based summarization—step one: Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)* (pp. 703–710). Austin, Texas.

Kupiec, Julian, Pedersen, Jan O., & Chen, Francine (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1995)* (pp. 68–73). Seattle, Washington.

Lease, Matthew, & Charniak, Eugene (2005). Parsing biomedical literature. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP'05)* (pp. 58–69). Jeju Island, Korea.

Lin, Chin-Yew, & Hovy, Eduard (2003). Automatic evaluation of summaries using *n*-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2003)* (pp. 71–78). Edmonton, Alberta.

Lin, Chin-Yew (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004) at ACL 2004*. Barcelona, Spain.

Ling, Xu, Jiang, Jing, He, Xin, Mei, Qiaozhu, Zhai, Chengxiang, & Schatz, Bruce (2006). Automatically generating gene summaries from biomedical literature. In *Pacific Symposium on Biocomputing, 11*, 40–51.

Lu, Zhiyong, Cohen K. Bretonnel, & Hunter, Lawrence (2006). Finding GeneRIFs via gene ontology annotations. In *Pacific Symposium on Biocomputing 11*, pp. 52–63.

Mani, Inderjeet, Gates, Barbara, & Bloedorn, Eric (1999). Improving summaries by revising them. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)* (pp. 558–565). College Park, Maryland.

Mani, Inderjeet, Klein, Gary, House, David, & Hirschman, Lynette (2002) SUMMAC: A text summarization evaluation. *Natural Language Engineering, 8*(1), 43–68.

McKeown, Kathleen R., Elhadad, Noemie, & Hatzivassiloglou, Vasileios (2003). Leveraging a common representation for personalized search and summarization in a medical digital library. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2003)*. (pp. 159–170). Houston, Texas.

Minel, Jean-Luc, Nugier, Sylvaine, & Piat, Gerald (1997). How to appreciate the quality of automatic text summarization? Examples of FAN and MLUCE protocols and their results on SERAPHIN. In *Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain.

Mizzaro, Stefano (1998). How many relevances in information retrieval? *Interacting With Computers, 10*(3), 305–322.

Morris, Andrew, Kasper, George, & Adams, Dennis (1992). The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research, 3*(1), 17–35.

Paice, Chris D. (1990). Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management, 26*(1), 171–186.

Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)* (pp. 311–318). Philadelphia, Pennsylvania.

Ruch, Patrick, Chichester, Christine, Cohen, Gilles, Coray, Giovanni, Ehrler, Frédéric, Ghorbel, Hatem, Müller, Henning, & Pallotta, Vincenzo (2003). Report on the TREC 2003 experiment: Genomic track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*. Gaithersburg, Maryland.

Salton, Gerard, Singhal, Amit, Mitra, Mandar, & Buckley, Chris (1997). Automatic text structuring and summarization. *Information Processing and Management, 33*(2), 193–207.

Smith, Lawrence H., Rindflesch, Thomas C., & Wilbur, W. John. (2005). The importance of the lexicon in tagging biological text. *Natural Language Engineering, 12*(2), 1–17.

Smucker, Mark, & Allan, James (2006). Find-Similar: Similarity browsing as a search tool. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)* (pp. 461–468). Seattle: Washington.

Turner, Jenine, & Charniak, Eugene (2005). Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)* (pp. 290–297). Ann Arbor, Michigan.

Wang, Ruichao, Stokes, Nicola, Doran, William, Newman, Eamonn, Carthy, Joe, & Dunnion, John (2005). Comparing Topiary-style approaches to headline generation. In *Lecture Notes in Computer Science: Advances in Information Retrieval: 27th European Conference on IR Research (ECIR 2005)*, Vol. 3408. Santiago de Compostela, Spain: Springer Berlin/Heidelberg.

Wilbur, W. John, & Coffee, Leona (1994). The effectiveness of document neighboring in search enhancement. *Information Processing and Management, 30*(2), 253–266.

Wilbur, W. John (2005). Modeling text retrieval in biomedicine. In Chen, Hsinchun, Fuller, Sherrilynne S., Friedman, Carol, & Hersh, William (Eds.), *Medical informatics: Knowledge management and data mining in biomedicine* (pp. 277–297). New York: Springer Science.

Zajic, David, Dorr, Bonnie J., & Schwartz, Richard (2004). BBN/UMD at DUC-2004: Topiary. In *Proceedings of the 2004 Document Understanding Conference (DUC 2004) at NLT/NAACL 2004*. Boston, Massachusetts.

Zajic, David, Dorr, Bonnie, Lin, Jimmy, & Schwartz, Richard (2007). Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management, Special Issue on Summarization*, in press.

Zajic, David M. (2007). *Multiple Alternative Sentence Compressions (MASC) as a tool for automatic summarization tasks*. Ph.D. thesis, College Park: University of Maryland.

Zhou, Liang, & Hovy, Eduard (2003). Headline summarization at ISI. In *Proceedings of the HLT-NAACL 2003 Text Summarization Workshop and Document Understanding Conference (DUC 2003)* (pp. 174–178). Edmonton, Alberta.