Chapter 27
Newness and Givenness of Information:
Automated Identification in Written Discourse

Philip M. McCarthy
The University of Memphis, USA

Zhiqiang Cai
The University of Memphis, USA

David Dufty
The Australian Bureau of Statistics, Australia

Danielle S. McNamara
Arizona State University, USA

Christian F. Hempelmann
Purdue University, USA

Arthur C. Graesser
The University of Memphis, USA

ABSTRACT

The identification of new versus given information within a text has been frequently investigated by researchers of language and discourse. Despite theoretical advances, an accurate computational method for assessing the degree to which a text contains new versus given information has not previously been implemented. This study discusses a variety of computational new/given systems and analyzes four typical expository and narrative texts against a widely accepted theory of new/given proposed by Prince (1981). Our findings suggest that a latent semantic analysis (LSA) based measure called span outperforms standard LSA in detecting both new and given information in text. Further, span outperforms standard LSA for distinguishing low versus high cohesion versions of text. Our results suggest that span may be a useful variable in a wide array of discourse analyses.

INTRODUCTION

One of the fundamental questions in discourse processing is how to differentiate new information from given information (Clark & Haviland, 1977; Haviland & Clark, 1974; Kennison & Gordon, 1997; Poesio & Vieira, 1998; Prince, 1981). Given information matches antecedent information in the text, discourse space, or common ground between speaker and listener (Clark, 1996), whereas new information expands on the body of given information. Differentiating new versus given information applies to written text

DOI: 10.4018/978-1-60960-741-8.ch027

as well as oral conversation. To better understand the relationship between given text and new text, consider the following exchange in a conversation. *Person One* says "I haven't seen much of Jerry lately." *Person Two* replies "Jerry has a new job." In Person Two's reply, "Jerry" has already been introduced into the conversation. That part of Person Two's speech act is given rather than new information because it has already been introduced into the discourse space. In contrast, "has a new job" is new information. For written text, consider the following passage from the online edition of the *Wall Street Journal* (01-05-2008):

Time Warner Inc. Chief Executive Jeff Bewkes pulled the trigger on his first major move to shake up the company, unveiling plans to spin off Time Warner Cable Inc. But investors gave the widely-telegraphed move a lukewarm reception and shifted their attention to the fate of the AOL unit.

In this example, the second sentence refers to several ideas that are mentioned in the previous sentence but also introduces much new information. For example, the sentence refers to "investors," and "lukewarm reception," both new pieces of information. On the other hand, "the widely telegraphed move" clearly refers to the spin-off plans described earlier. This component of the sentence is not *new* because it has already been *given* to the reader. Interestingly, the content words do not overlap between the two constituents. This observation illustrates that the challenge of computing given information is much more complex than merely computing overlap words between an incoming sentence and the prior discourse context.

The importance of the new/given distinction is widely accepted, but there is not a uniform consensus on what counts as new versus given information. Does given information refer only to explicit antecedent information or can it refer to inferences suggested by the text? Does given information include shared knowledge of people in a community (e.g., the president of a country) or is it necessary to introduce given information in the verbal discourse or physical context of a particular spoken conversation? If we attempted to program a computer to compute new versus given information, what sort of algorithms would be adequate? Is it even possible to devise a complete and reliable algorithm? If not, then it will always be necessary for discourse processing researchers to annotate new versus given information by hand.

The present study examines some automated algorithms for computing new versus given information in printed text as well as conversational interaction. Some algorithms will be standard components developed in the field of computational linguistics (Jurafsky & Martin, 2008), whereas others will be statistical algorithms developed in cognitive science. Most notably, the primary statistical algorithm in this study, span, is a variant of Latent Semantic Analysis (LSA, Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). LSA is the core component of a number of automated essay graders that can evaluate essays as reliably as expert human graders (Burstein, 2003; Landauer, Laham, & Foltz, 2003). LSA has also been used for a variety of other applications, such as information retrieval (Deerwester, Dumais, Furnas, Landauer, & Harshman 1990), automated tutoring systems (Graesser, Lu, et al. 2004; McNamara, Levinstein, & Boonthum, 2004), evaluation of text coherence (Foltz, Kintsch, & Landauer,1998; Graesser, Jeon, Yang, & Cai, 2007; McNamara, Cai, & Louwerse, 2007), text type identification (McCarthy, Briner, Rus, & McNamara, 2007), and assessments of reading comprehension (Millis et al., 2004).

The LSA technique (see Chapter 9 for more details) requires a corpus analysis in which occurrences of all words in the corpus are recorded in a very large word-by-document matrix. This matrix is then reduced in size using a statistical compression technique called singular value decomposition. The resulting smaller matrix is referred to as the *LSA space*. The similarity of two words (or sentences, paragraphs, or entire texts) is computed by the similarity of their vectors in the LSA space. One virtue of LSA is that it can