Transitivity and Foregrounding in News Articles: experiments in information retrieval and automatic summarising

Roderick Kay and Ruth Aylett
Information Technology Institute
University of Salford
Manchester M5 4WT
United Kingdom
{rnk,R.Aylett}@iti.salford.ac.uk

Abstract

This paper describes an on-going study which applies the concept of transitivity to news discourse for text processing tasks. The complex notion of transitivity is defined and the relationship between transitivity and information foregrounding is explained. A sample corpus of news articles has been coded for transitivity. The corpus is being used in two text processing experiments.

1 Introduction

The basic hypothesis of this study is that the degree of transitivity associated with a clause indicates the level of importance of a clause in a narrative text. For this assumption to form the basis of a practical implementation, transitivity must be objectively defined and the definition must be able to be processed automatically.

The notion of transitivity clearly has many implications for text processing, in particular information retrieval and automatic summarising, because it can be used to grade information in a document according to importance. In an information retrieval context, it means that a transitivity index could influence a decision about the relevance of a document to a query. In automatic summarising, it means that less important information could be *sieved* according to transitivity, leaving only the most important information to form the basis of a summary.

News discourse was chosen because it is narrative based and therefore broadly applicable to the notion of transitivity. There has also been extensive research in the structural characteristics of this text type (Duszak, 1995) (Kay and Aylett, 1994) (Bell, 1991). However, the study poses a challenge in the sense that the notion of transitivity has previously been exemplified with relatively simple sentences presenting action sequences. A central question is how well the concept can be transferred to a domain which, although narrative based, diverges into commentary and analysis.

2 Definition of Transitivity

Transitivity is usually considered to be a property of an entire clause (Hopper and Thompson, 1980). It is, broadly, the notion that an activity is transferred from an agent to a patient. It is therefore inherently linked with a clause containing two participants in which an action is highly effective.

The concept of transitivity has been defined in terms of the following parameters:

A. Participants	h	2 or more participants
	l	1 participant
B. Kinesis	h	action
	l	non-action
C. Aspect	h	telic
	l	atelic
D. Punctuality	h	punctual
	l	non-punctual
E. Volitionality	h	volitional
	l	non-volitional
F. Affirmation	h	affirmative
	l	negative
G. Mode	h	realis
	l	irrealis
H. Agency	h	A high in potency
	l	A low in potency
I. Affectedness of O	h	O totally affected
	l	O not affected
J. Individuation of O	h	O highly individuated
	l	O non-individuated

h=high, l=low

Each component of transitivity contributes to the overall effectiveness or 'intensity' with which an action is transferred from one participant to another.

A. There must be at least two participants for an action to be transferred.
B. Transferable actions can be contrasted with non-transferable states, e.g. he pushed her; he thought about her.
C. An action is either wholly or partially completed according to whether it is telic or atelic, e.g. I played the piano; I am playing the piano.
D. Punctual actions have no transitional phase between start and end point, having a greater effect

on their patients, e.g. he kicked the door; he opened the door.
E. An action is more effective if it is volitional, e.g. he bought the present; he forgot the present.
F. An affirmative action has greater transitivity than a negative action, e.g. he called the boy; he didn't call the boy.
G. An action which is realis (occurring in the real world) is more effective than an action which is irrealis (occurring in a non-real contingency world), e.g. they attacked the enemy; they might attack the enemy.
H. Participants high in agency transfer an action more effectively than participants low in agency, e.g. he shocked me; the price shocked me.
I. A patient is wholly or partially affected, e.g. I washed the dishes; I washed some of the dishes.
J. Individuation refers to the distinctiveness of the object from the agent and of the object from its own background. The following properties contribute to the individuation of an object.

INDIVIDUATED	NON-INDIVIDUATED
proper	common
human, animate	inanimate
concrete	abstract
singular	plural
count	mass
referential, definite	non-referential

Based on these components, clauses can be classified as more or less transitive. In English, as a whole, transitivity is indicated by a cluster of features associated with a clause.

The concept of foreground and background information is based on the idea that in narrative discourse some parts are more essential than others. Certain sections of a narrative are crucially linked with the temporal sequence of events which form the backbone of a text. This material is normally foregrounded. In contrast, the contextual information relating to characters and environment is backgrounded.

3 Transitivity and Text Processing

The relationship between transitivity and foregrounding has potential for text processing, in particular, information retrieval and automatic summarising. If it is possible to identify which clauses are central to a text, the information can be used to contribute to a relevance assessment or as the basis for a derived summary.

3.1 Information Retrieval

The standard model of text retrieval is based on the identification of matching query/document terms which are weighted according to their distribution throughout a text database. This model has also been enhanced by a number of linguistic techniques: expansion of query/document terms according to thesaurus relations, synonyms, etc.

The proposal for this study is to code matching query/document terms for the transitivity value of the clause in which they occur, as a starting point for producing comparative term weights based on linguistic features. Terms which are less central to a discourse will, on this basis, be given lower scores because they occur in low transitivity clauses. The net result will be to produce a document ranking order which more closely represents the importance of the documents to a user. There is also potential for producing a transitivity index for an entire document as well as for individual clauses so that this measure could also feature in a relevance assessment.

3.2 Automatic Summarising

The fundamental task in automatic summarising is to identify the most important sections of a text so that these can be extracted and possibly modified to provide a summary. The notion of transitivity provides a measure against which clauses can be scored. The highest scoring clauses, either above a threshold value or on a comparative basis, can then be identified as the basic clauses of a summary. These can either be extracted *raw* or in context with pronominal references resolved and any logical antecedents included. A previous study in this area (Decker, 1985) extracted clauses and sentences on the basis of syntactic patterns which broadly correlate with certain features of transitivity. The present study focuses on the semantic features of transitivity rather than associated syntax.

4 Experimental Procedure

The feasibility of using transitivity as a tool in text processing will be assessed by two experiments using the same corpus. Clauses in the corpus must be hand-coded for transitivity. The difficulties encountered in this process will determine the basis for future automation. For the information retrieval task, only the clauses containing query/document matching terms will be coded for transitivity. For the automatic summarising experiment all sentences within a text will be coded.

For the information retrieval experiment, ten queries are put to a newspaper database: a demonstration system running on WAIS (Wide Area Information Server), carrying two weeks of articles from the Times newspaper from 1993 and 1994. The results of the queries are downloaded in their initial ranked order (ranked by a host ranking algorithm) and re-ranked by a serial batch processor written in C++. The processor identifies the transitivity features associated with each matching clause and produces a ranked output of documents based on the weights assigned to each clause in which the search terms occur. The weights assigned to each clause are

numerically equivalent to the number of transitivity features associated with each clause. The total transitivity weight for an entire document is the sum of clause weights normalised by document length.

The output dataset consists of a total of 185 news articles, an average of 18.5 per batch. Each set of articles is ranked by volunteers. The articles are ranked for their degree of relevance to a query in two ways: on a scale of one to ten; and comparatively, by the degree of relevance of an article against all other articles. All terms are treated as equal so that discrimination between documents is based purely on accumulative transitivity scores. The performance of the ranking technique is evaluated according to two precision measures: the Spearman rank correlation coefficient (rho) and the CRE (Coefficient of Ranking Effectiveness) (Noreault et al. , 1977).

For the automatic summarising experiment, ten articles are taken from the corpus at random. Summaries are produced by extracting clauses according to transitivity scores. In the initial implementation, transitivity scores will be equal to the number of transitivity features associated with the main clause of each sentence. The selection of sentences for a summary will be based, initially, on comparative transitivity scores and a reduction factor which will determine the number of sentences selected based on the length of a document.

Summaries will be analysed and assessed by volunteers for coverage, in terms of the original text, and comprehensibility as a separate text. The summaries will be compared against summaries of the same texts compiled by the syntactic technique mentioned previously and also against summaries consisting of the first paragraph of each news article.

The study is currently at the end of the coding stage for the information retrieval experiment.

References

A. Bell. 1991. The language of news media. *Basil Blackwell, Oxford*

N. Decker. 1985. The use of syntactic clues in discourse processing. *Proceedings of the 23rd meeting of the ACL*, pages 315–323.

A. Duszak. 1995. On variation in news-text prototypes: some evidence from English, Polish, and German. *Discourse Processes*, 19: 465–483.

G. Green. 1979. Organization, goals and comprehensibility in narratives: news writing, a case study. *Technical Report 132*, The Centre for the study of Reading, University of Illinois at Urbana-Champaign.

R. Kay, R. Aylett. 1994. A text grammar for news reports. *Working papers in Language and Linguistics, No 2*, European Studies Research Institute, University of Salford.

P. Hopper, S. Thompson. 1980. Transitivity in grammar in discourse. *Language*, 56: 251–299.

T. Noreault, M. Koll, M. McGill. 1977. Automatic ranked output from Boolean searches in SIRE. *Journal of the American Society for Information Science*, 27(6): 333-339.