Description of title sentence attributes used in classification

James M. Eales, George Demetriou, Robert Stevens

1 Attribute details

Table 1 lists some of the attributes calculated for each article title, the remaining attributes are described in section 1.1.

1.1 Additional attributes

1.1.1 Typed dependencies

We also include a set of values which provide the proportion of dependency links between phrases in the title for each of the standard Stanford parser dependency types (see `http://nlp.stanford.edu/software/dependencies_manual.pdf`). The names of these attributes follow the pattern 'prop' *short dependency name (first letter capitalised)* 'Dep'. Examples are 'propNnDep', 'propDobjDep' and 'propPartmodDep'.

1.1.2 Part-of-speech tag n-grams (POS-grams)

The sequence of part-of-speech tags, from each title, are first shortened to a base form (e.g. NNP becomes NN and VBZ becomes VB) and then split into bigrams and trigrams and these are screened for the presence of 717 unique (shortened) bi- and trigrams that occur commonly in titles. Therefore the values for the POS-gram attributes are numerical integers that show the number of times a given POS-gram appears in the title.

Table 1: Table of article title attributes and their description

Attribute name	**Attribute type**	**Description**
goodTitleOrNot	Binominal (Class)	('good', 'bad')
posTagOfFirstToken	Nominal	Part-of-speech tag for first token
posTagOfSecondToken	Nominal	Part-of-speech tag for second token
posTagOfThirdToken	Nominal	Part-of-speech tag for third token
containsVerb	Binominal	Does the title contain a verb
containsColon	Binominal	Does the title contain a colon
containsQuestionMark	Binominal	Does the title contain a question mark
containsHyphen	Binominal	Does the title contain a hyphen
lengthInTokens	Numerical	Number of tokens (similar to number of words) in the title
lengthInCharacters	Numerical	Number of text characters in the title
numberOfSentences	Numerical	Number of sentences in title, determined by OpenNLP sentence splitter
numberOfNamedEntities	Numerical	Total number of named entities found by four different named entity recognition tools
numberOfProteinEntities	Numerical	Number of unique proteins found by whatizit pipeline 'whatizitSwissprot'
numberOfSmallMoleculeEntities	Numerical	Number of unique small molecules found by whatizit pipeline 'whatizitChebiDict'
numberOfDiseaseEntities	Numerical	Number of unique diseases found by whatizit pipeline 'whatizitDisease'
numberOfCellTypeEntities	Numerical	Number of unique cell types found by uniman cell types Web Service
proportionOfNouns	Numerical	Proportion of title tokens that are nouns (determined by OpenNLP POS tagger)
proportionOfVerbs	Numerical	Proportion of title tokens that are verbs (determined by OpenNLP POS tagger)
proportionOfAdjectives	Numerical	Proportion of title tokens that are adjectives (determined by OpenNLP POS tagger)
proportionOfHedges	Numerical	Proportion of title tokens that are hedge words (e.g. 'may', 'potential', 'could')
proportionOfAdjectiveChunks	Numerical	Proportion of adjective phrases in title (determined by OpenNLP chunker)
proportionOfAdverbChunks	Numerical	Proportion of adverb phrases in title (determined by OpenNLP chunker)
proportionOfNounChunks	Numerical	Proportion of noun phrases in title (determined by OpenNLP chunker)
proportionOfPrepositionChunks	Numerical	Proportion of preposition phrases in title (determined by OpenNLP chunker)
proportionOfVerbChunks	Numerical	Proportion of verb phrases in title (determined by OpenNLP chunker)
verbOne	Nominal	The first verb (stemmed) found in the ordered list of title tokens (stemmed using Stanford parser morphological analyser)
verbTwo	Nominal	The second verb (stemmed) found in the ordered list of title tokens (stemmed using Stanford Parser morphological analyser)
verbThree	Nominal	The third verb (stemmed) found in the ordered list of title tokens (stemmed using Stanford Parser morphological analyser)
numberOfTypedDependencies	Numeric	Number of typed dependencies (links between phrases) in the title, value derived using Stanford parser
phraseStructureMaxDepth	Numeric	Maximal depth of phrase structure in the title, calculated by Stanford parser