Preliminary study for feature selection

Preliminary study. For this study, we collected 500 documents by querying common search engines such as Bing with 5 keywords from the domains of health, society and economics. The retrieved documents have been segmented into around 4000 paragraphs. Then, five experts assessed for each paragraph whether it contains an argument. For the 193 paragraphs that contain arguments, we then calculated the values for a set of features that are commonly used in argumentation mining and NLP [1]. For this training dataset, we selected the features that turned out to be good predictors for paragraphs that contain arguments. Following a best-first feature selection strategy [2], we ended up with the following set of features.

Lexical features. Lexical features deal with the words or vocabulary of a language. The following lexical information of a paragraph turned out to be useful to classify its argumentative nature.

Thematic words: the most frequent words (ignoring common stop-words, such as connectives and articles) in a document and thus paragraph are considered to be thematic words. Selecting a small number of thematic words that are particularly relevant to the query keyword, this feature is defined as the frequency counts of thematic words. The intuition of this feature is that a paragraph containing the most important keywords should express the argumentative point of view of the writer.

Example 1 *Consider the following article regarding vaccine* [1]*, there are various words that appear in high frequency in the documents such as vaccine, shots, disease, autism, health. This means these words discuss the theme of the documents. Among the paragraphs in this document, the following paragraph contains many thematic words, which shows that it may contains an argument:*

(S1) Yes. (S2) Vaccines are safe. (S3) In fact, experts including American Academy of Pediatrics, the Institute of Medicine, and the World Health Organization agree that vaccines are even safer than vitamins. (S4) Millions of children and adults are vaccinated every yearsafely. (S5) Thousands of people take part in clinical trials to test a vaccine before it is licensed by the Food and Drug Administration (FDA). (S6) After its licensed, the Vaccine Adverse Events Reporting System (VAERS) helps track any health effect that happens hours, days, weeks, or even months later. (S7) Anyone can report a possible side-effect so that it can be studied. (S8) This monitoring helps ensure vaccines are safe. (S9) To learn more about vaccine safety from the Centers for Disease Control and Prevention, visit the CDC vaccine safety page.

Segment S_2 is likely to be a claim as the segment is the second sentence in the paragraph and the keyword vaccine is the subject. Segment S_3 may be an evidence as it contains evidence-related word (fact) and names (American Academy of Pediatrics,

[1] http://www.whyichoose.org/vaccinesafety.html

Institute of Medicine, World Health Organization). Segment S_4 and S_5 can also be evidence as they contains numbers (millions, thousands).

Evidence-related words: A paragraph containing many evidence-related words, such as numbers, citations, and cue phrases (e.g., 'because of') is likely to contain arguments. This feature, thus, is defined as the occurrence count of such evidence indicators.

Example 2 *(S1) Processed foods destroy your mind. (S2) If you suffer from chronic bouts of brain "fog," or have difficulty concentrating and thinking normally, chances are your diet has something to do with it. (S3) And a recent study out of Oxford University lends credence to this possibility, having found that junk food consumption can cause people to become angry and irritable.(S4) Nutrient-dense whole foods, on the other hand, can help level out your mood, sustain your energy levels, and leave you feeling calmer and more collected.*

This paragraph may contain an argument as an evidence-related keyword (study) is available in segment S_3 of the paragraph. In addition, the keyword processed foods appears as a subject of the first sentence, which shows that this sentence is likely to be a claim.

Prototypical words: Prototypical words are lexical expressions used to formulate arguments, for instance, 'argue' or 'believe'. As part of our preliminary study, we learnt a list of 97 prototypical words from our training data. Each prototypical word is associated with a weight, indicating the likelihood of a paragraph containing the given word to include an argument. This weight is determined based on the relative frequency of the word in the training data and in the overall set of documents. Again, the respective feature is the occurrence count of the prototypical words, normalized by their respective weight.

Example 3 *(S1) Believe it or not, almost all the food that you eat, even the foods made from scratch," have actually been processed. (S2) According to an article published in the journal, Advances in Nutrition, any food that has been subject to washing, cleaning, milling, cutting, chopping, heating, pasteurizing, blanching, cooking, canning, freezing, mixing, and packaging that alter the food from its natural state is considered a processed food.*

The above paragraph contains an argument as segment S_1 which contains a prototypical word (believe). This prototypical word signifies that it may be a claim. In addition, segment S_2 provides information coming from a reliable source which is the Advances in Nutrition *journal*

Syntactical features. Syntactical features refer to the text structure and capture local relations between words within a sentence. Our preliminary study identified the following syntactical features:

- *Part-of-speech*: Words in a sentence may be classified into different parts-of-speech, such as nouns, verbs and adjectives. Intuitively, this feature exploits the fact that, if the query keyword appears to be a subject or object of a sentence, it is likely that the sentence is a relevant claim or evidence. Technically, this feature is defined as

the occurrence count of the query keyword in a paragraph either as a subject or as an object.

Example 4 *Consider segment S_2 of Example 1 and segment S_1 of Example 2, these segments have the keywords (vaccine and processed foods) appearing as the subject of a sentence. This sentence may be a claim, hence, the paragraph may contain an argument.*

- *Parse structure*: A sentence can be parsed into a tree-like structure that captures syntactical relations between the phrases within that sentence. In the parse structure, the relation between the phrases is determined by the *head word*, which is an indicator of the topic being mentioned. We use the appearance of the query keyword as the head word of phrases in a paragraph as a classification feature indicating whether the paragraph contains arguments.

For illustration, we consider the following paragraph that is relevant for the topic of 'processed food' discussed above.

Example 5 *Furthermore, the engineering behind processed food makes it virtually addictive. A 2009 study by the Scripps Research Institute indicates that overconsumption of processed food triggers addiction-like neuroaddictive responses in the brain, making it harder to trigger the release of dopamine. In other words the more processed food we eat, the more we need to give us pleasure; thus the report suggests that the same mechanisms underlie drug addiction and obesity.*

This paragraph illustrates various of the features that hint at arguments, see also Table 1. Examples include evidence-related words, such as 'study' and 'report' and prototypical words, such as 'indicate'. In addition, thematic words, e.g., 'addictive' and 'addition' render the paragraph important for argumentation mining. Further, the query keyword 'processed foods' appears several times. In the first sentence, part-of-speech tagging identifies the keyword as the subject of a sentence. In the second sentence, the keyword appears as the head word for the phrase 'overconsumption of processed food', which is yet another indicator that the paragraph contains an argument.

Table 1: Features and example values

Feature types	**Example values for keyword 'processed foods'**
Thematic words	additive, addiction
Evidence-related words	because of, by the fact that
Prototypical words	argue, claim, believe
Keyword as subject or object	the engineering behind processed food makes [...]
Keyword as head word	overconsumption of processed foods

References

[1] R. Palau et al. "Argumentation mining: the detection, classification and structure of arguments in text". In: *ICAIL*. 2009.

[2] L. Xu et al. "Best first strategy for feature selection". In: *ICPR*. 1988.