===== Introduction =====

This corpus contains sentences from the abstract and introduction of 30 scientific articles that have been annotated (i.e. labeled or tagged) according to a modified version of the Argumentative Zones annotation scheme [1]. These 30 scientific articles come from three different domains:

1. PLoS Computational Biology (PLOS)
2. The machine learning repository on arXiv (ARXIV)
3. The psychology journal Judgment and Decision Making (JDM)

There are 10 articles from each domain. In addition to the labeled data, this corpus also contains a corresponding set of unlabeled articles. These unlabeled articles also come from PLOS, ARXIV, and JDM. There are 300 unlabeled articles from each domain (again, only the sentences from the abstract and introduction). These unlabeled articles can be used for unsupervised or semi-supervised approaches to sentence classification which rely on a small set of labeled data and a larger set of unlabeled data.

===== Description of the annotation scheme =====

Argumentative Zones (AZ) is a scheme for classifying (i.e. annotating, labeling, or tagging) sentences according to function. See [1] for an introduction to AZ. There are seven labels in the original AZ scheme:

1. AIM: "A specific research goal of the current paper"
2. TEXTUAL: "Statements about section structure"
3. OWN: "(Neutral) description of own work presented in current paper"
4. BACKGROUND: "Generally accepted scientific background"
5. CONTRAST: "Statements of comparison with or contrast to other work; weaknesses of other work"
6. BASIS: "Statements of agreement with other work or continuation of other work"
7. OTHER: "(Neutral) description of other researchers' work"

These descriptions were taken from Table 1 in [1]. The annotation scheme used in this corpus is derived from AZ and includes five labels:

1. AIM
2. OWN
3. CONTRAST
4. BASIS
5. MISCELLANEOUS

The labels AIM, OWN, CONTRAST, and BASIS correspond to the labels in the original AZ scheme. The label MISCELLANEOUS is a combination of BACKGROUND and OTHER since we found, in practice, that it was difficult to distinguish between the two labels. The label TEXTUAL from the original AZ scheme has been removed.

===== Directory structure =====

This directory contains 3 subdirectories:

1. "unlabeled_articles"

The directory "unlabeled_articles" contains 3 directories -- one for each of the 3 domains -- entitled "arxiv_unlabeled", "jdm_unlabeled", and "plos_unlabeled" respectively. Each of these "domain" directories contains 300 unlabeled articles. Each article is identified by an unique integer article identification number.

2. "labeled_articles"

This subdirectory contains 90 plaintext files: 3 files containing 3 independent annotations for each of the 30 selected scientific articles. The filenames have the following format:

[DOMAIN]_annotate[1-10]_[ARTICLE_ID]_[ANNOTATOR].txt

where

DOMAIN is either "arxiv", "jdm", or "plos"
ARTICLE_ID is the unique integer article identification number for the article
ANNOTATOR is either 1, 2, or 3 corresponding to the first, second, or third annotator respectively

For example, "jdm_annotate10_210_1.txt" contains the annotations provided by the first annotator for the JDM article with identification number 210. Similarly, "jdm_annotate10_210_3.txt" contains the annotations provided by the third annotator for that same article.

3. "word_lists"

This directory contains one plaintext file for each of the 4 labels AIM, OWN, CONTRAST, BASIS. Each plaintext file lists the indicator words for the corresponding label. That is, those words that are semantically indicative of the label. For example, the indicator words for the label CONTRAST include the words "not", "difficult", "limited", "however", "assume", "only", "many", "although", "typically", "directly". The indicator words for the label OWN include the words "we", "section", "our", "this", "results", "used", "new". These word lists were subjectively created from the labeled data.

This directory also contains a stopwords file. This stopwords file contains stopwords that are not likely to be important for the task of sentence classification. For example, the words "how", "show", "we", and "our" are often considered stopwords. However, for sentence classification these words are actually strong features that we do not want to ignore. The stopwords file contains a set of 25 stopwords that are not likely to be strong features for this task and thus can be safely removed.

===== Annotation procedure =====

Each of the 30 scientific articles was labeled by 3 independent annotators. Two of those three annotators were not likely to be experts in the scientific domain

(i.e. computational biology, machine learning, or psychology). Nor were they experts in the task of annotating sentences. The third annotator, in contrast, did have training in annotating sentences according to AZ. The annotations from this annotator are always given the ANNOTATOR id of 3 (see the filename format description above).

Each participant received a directory containing a text version of the article to be labeled, a pdf version of the article to be labeled, a 7-page instruction manual, and three examples of already-labeled articles. Participants were told to read the instruction manual and encouraged to look at the three examples before starting the task. The instruction manual was taken and modified from the instructions written by Teufel et al. [2].

Since the participants were not likely to be experts in the scientific domain nor experts at annotating sentences and they received no training beyond the instruction manual (and instructions received via email regarding any clarifying questions the participants may have had), we used two methods to ensure the quality of the annotations. First, the last sentence of the instruction manual asked the participants to respond by email with the time of day. We used this to assess who read the entire instruction manual. Second, we inspected each labeled article and checked that any sentences with the words "we" or "our" were labeled with either AIM or OWN as specified in the instruction manual. We also checked for general understanding of the task, e.g. one participant marked all sentences that did not discuss the author's own work as either BASE or CONTRAST -- they marked no MISCELLANEOUS sentences. Any participant that failed to pass these two quality checks was not asked to participate any further and the labels they provided were not used. Occasionally, a participant demonstrated a desire to correct their mistakes and learn what they had done wrong. In these instances, we provided feedback and allowed the participant to try a second article. Those participants that did pass the quality checks were paid a $15 dollar Amazon gift card and were allowed to do additional articles for $5 each (or $7 for PLOS articles) up to a total of a $50 Amazon gift card.

We took as ground truth for each sentence, the majority label. That is, the label assigned by a majority (i.e. at least two) of the three annotators. For 2 sentences, there was no majority label in which case the label provided by annotator three (the "expert" annotator) was used.

A pdf of the instructions given to the annotators is provided along with this data set. The instruction manual provides a clearer understanding of the semantics of each of the labels. The pdf is called "Instructions_for_SentenceAnnotation.pdf"

===== References =====

[1] S. Teufel and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics, 28(4):409-445, 2002.

[2] S. Teufel. Argumentative zoning: information extraction from scientific text. PhD thesis, School of Informatics, University of Edinburgh, 1999.