Feature Engineering for SpaceEval

Seth Dworman

17 December 2014

1.0 Task Description

ISO-Space is a rich annotation markup language for capturing spatial and motion based information in natural language text. One of the overarching goals in creating the ISO-Space gold standard is to provide supervised data for machine learning of annotation language. That is, given unstructured natural language text, an algorithm should generate the ISO-Space markup automatically. This task by itself is rather difficult due to the diversity of spatial language, the richness of the ISO-Space specification, and the quality, nature, and distributions of the gold standard data. Simple bag-of-word approaches are not sufficient as spatial language is itself relatively sparse even in the gold standard. In addition, identification of spatial mentions (=text extents which should be classified into an ISO-Space tag) is not sufficient for generating the full ISO-Space markup, as each tag itself has its own attribute values, some of which are not trivial and must also be inferred by an algorithm. Lastly, once all mentions are determined, it is necessary to link them correctly via the ISO-Space links.

2.0 Feature Engineering

We propose that bag-of-word approaches will not suffice for learning the ISO-Space language or developing a suitable generalized algorithm. To get a reasonable algorithm, we suggest several features to supplement a bag-of-words approach. We follow the **statistical language model representation hypothesis** (known as LMRH), which claims that features that are informative for one NLP task are also informative on others (Huang, Ahuja, Downey, Yang, Guo, and Yates 2014). The features we choose to select for are part-of-speech tags (POS tags), named-entity recognition tags (NER tags), and semantic labels (Sparser labels) from Sparser, a handwritten shallow semantic based parser (McDonald).

3.0 Training Instances

The ISO-Space specification works at the sentential level, with only possibly METALINKs being intersential relations, as they are essentially coreference links limited to spatial mentions. We will ignore this for now and assume that an algorithm generating ISO-Space markup would operate with a single sentence as its input (with larger documents simply iterated over). Below we have reproduced an example sentence, *I have taken a boat up Amazon, crossed dirt roads over 15,000 ft Andean passes, luckily escaped bandits in southern Mexico, felt the wind and rain of of Patagonia, and dodged buses in the capital cities of nearly every country.*, already tokenized for us from the gold standard set. The column named **Label** represents the token's ISO-Space tag value, with NONE indicating it is not a spatial mention. We interpret **Label** as the true class label y we wish to predict given a token. This example is taken from `Tokenized/RFC/BuenosAires.xml`.

Token	Label
I	SPATIAL_ENTITY
have	NONE
taken	MOTION
a	MOTION_SIGNAL
boat	MOTION_SIGNAL
up	MOTION_SIGNAL
the	NONE
Amazon	PATH
,	NONE
crossed	MOTION
dirt	NONE
roads	PATH
over	MOTION_SIGNAL
15,000	MEASURE
ft	MEASURE
Andean	NONE
passes	PATH
,	NONE
luckily	NONE
escaped	NONMOTION_EVENT
bandits	NONE
in	SPATIAL_SIGNAL

Token	Label
southern	NONE
Mexico	PLACE
,	NONE
felt	NONE
the	NONE
wind	SPATIAL_ENTITY
and	NONE
rain	SPATIAL_ENTITY
of	SPATIAL_SIGNAL
Patagonia	PLACE
,	NONE
and	NONE
dodged	MOTION
buses	SPATIAL_ENTITY
in	SPATIAL_SIGNAL
the	NONE
capital	NONE
cities	PLACE
of	SPATIAL_SIGNAL
nearly	NONE
every	NONE
country	PLACE
.	NONE

We consider each token from a sentence to be a possible spatial mention, and augment a simple bag-of-words approach with POS tags, NER tags, and Sparser labels. POS tags are relatively straightforward and come from Stanford NLP's POS tagger (Toutanova, Klein, Manning, and Singer 2003), which achieves a state-of-the-art performance using the Penn TreeBank POS labels. NER tags come from a small set of labels : `O(utside)`, `person`, `organization`, and `location`. We also make use of Stanford NLP's NER tagger (Finkel, Grenager, and Manning 2005). Our intuition is that tokens which are identified as a NER label are more likely to be involved in some kind of spatial relation (especially NER `location`), given that the domain of the gold standard is travel blogs. Finally, we use Sparser edge labels, which come from Sparser, a shallow semantic parser (McDonald). Sparser's labels are semantics based and form an open set. We believe that Sparser's labels can be useful for finding spatial mentions which are places, paths, measurements, and spatial signals, whose correct identification can bootstrap finding other spatial mentions (e.g. spatial named entities). Below we reproduce part of the original sentence (the first two clauses), augmented with POS, NER, and primary Sparser labels. Recall that the **Label** column should be interpreted as the class label y we wish to predict given a token.

Token	POS	NER	Sparser	Label
I	PRP	O	SINGLE-CAPITALIZED-LETTER	SPATIAL_ENTITY
have	VBP	O	HAVE	NONE
taken	VBN	O	None	MOTION
a	DT	O	INDIVIDUAL	MOTION_SIGNAL
boat	NN	O	INDIVIDUAL	MOTION_SIGNAL
up	RP	O	UP	MOTION_SIGNAL
the	DT	O	NAME	NONE
Amazon	NNP	LOCATION	NAME	PATH
,	None	O	None	None
crossed	VBD	O	CROSS	MOTION
dirt	NN	O	DIRT	NONE
roads	NNS	O	PATH-TYPE	PATH
over	IN	O	OVER	MOTION_SIGNAL
15,000	CD	O	None	MEASURE
ft	JJ	O	None	MEASURE
Andean	JJ	O	NAME	NONE
passes	NNS	O	PASS-KIND	PATH
,	None	O	None	None
luckily	RB	O	LUCKILY	NONE
escaped	VBD	O	ESCAPE-EVENT	NONMOTION_EVENT
bandits	NNS	O	BANDIT	NONE
in	IN	O	IN	SPATIAL_SIGNAL
southern	JJ	O	DIRECTION	NONE
Mexico	NNP	LOCATION	COUNTRY	PLACE

We note that Sparser also provides other labels besides the primary label shown; in fact all the field values of the Sparser `Edge` data structure are available as features, though some of these may not turn out to be useful.