**An ISO-Space Measure Extraction System**

Christopher Haberland

<https://github.com/argideritzalpea/575-Project>

**Introduction**

Enabling NLU systems to identify and infer relationships among spatial natural language in utterances and text would have widespread applicability for human-facing interfaces in the domains of robotics, GIS (geographic information systems), and virtual reality. Systems that correctly identify and interpret spatial language could allow for natural language commands to suffice for interacting with systems that primarily operate in or represent natural environments. SemEval 2015 Task 8 (SpaceEval) was a shared task that attempted to advance the state-of-the-art methods to bridge the gap between humans and machines with respect to spatial language understanding. This task was the first shared task to require participants to employ the ISO-Space annotation scheme (Pustejovsky 2012), which allows greater specificity and relational detail compared to the annotation scheme required for the spatial language-related SemEval task in 2013 (Pustejovsky 2015). While the 2013 task focused on the creation of systems for spatial language identification, the 2015 principally prompted participants to establish the linkages between entities and other language containing spatial information, such as phrases denoting orientation, relative distance, and verbs involving movement.

In the 2015 SpaceEval task, three teams competed against a baseline system created by the task authors to:

1. Annotate and identify spans containing information relevant to spatial information.
2. Identify relationships and linkages between entities and other spatial language to identify the spatial semantics of a passage.

The ISO-Space annotation scheme divides the semantics of spatial information contained in texts in several different labels, with each containing several different attributes to further define the nature of the spatial information inherent in the text. The training set in the 2015 SpaceEval task consisted of 55 trip narratives and was labeled according to the ISO-Space annotation scheme to be used as the gold standard. 14 additional narratives were annotated according to the ISO-Space scheme to be used as a test set. Participant systems were judged on their recall, precision, and F1-scores for predicting spatial language spans when compared to the gold standard. Three evaluation configurations were developed that allowed participants to maximize system performance according to span identification of specific sets of label types in ISO-Space. None of the three teams had the highest F1-scores for all three configurations. Partial span matching performance was not reported.

The purpose of this paper is to explore the label extraction task within the ISO-Space annotation environment. Specifically, I seek to create and test a rule-based system using WordNet to extract ‘measure’ labels from the SpaceEval trip narrative data. The motivation for this is to reduce the error-propagation problem inherent in creating systems that dually identify spatial language, entities, and the relationships and linkages between them. Properly identified candidate text spans will provide greater predictive power to a model that attempts to draw inferences about the spatial relationships between different text spans. This was witnessed in the SpaceEval task, as all systems performed much better using gold standard labels as compared to data labeled by the system itself.

**Prior Work**

No prior work has been published on specifically extracting “measure” labels within ISO-Space. The 2015 SpaceEval task did not address “measure” labels or “Mlink” classification in its prompts. Although the task prompted participants to identify and annotate all other relations and entities in the ISO-Space ontology, the task instructions specifically omit ‘measures’ and ‘MLinks’ (relations between ‘measures’ and other entities) from the task. The participant systems in the SpaceEval 2015 task are primarily interested with task of joining the spatial linkages between various entities in a text as opposed to annotating the entities and spatial markers according to individual spans of text. Nevertheless, the participant systems in SpaceEval 2015 represent interesting approaches to solving tasks that are dependent upon the type of spatial role labeling that is treated by the system proposed in this paper.

The systems submitted in the SpaceEval task were focused on extracting ‘MoveLinks’, ‘OLinks’, and ‘QLinks’ – not ‘MLinks’. One team from the University of Texas – Dallas (UTD) employed an ensemble approach of 31 features from among lexical, grammatical, semantic, positional, distance, entity attribute, and entity roles types to train an SVM classifier (D’Souza and Ng 2015). After presenting the initial model, the UTD team expanded their feature set to include expandable parse trees and employed sieve-based approach that increased their predictive accuracy of ‘MoveLinks’ with respect to their baseline model by 3% when using gold standard labels as input (45% to 48%). Interestingly, their highest-achieving system only attained 23.3% recall and an F1-score of 20.3 when using *extracted* spatial tags as input as opposed to gold standard labels. The salient drop in performance when using extracted labels is a motivating fact that highlights the importance of correct extraction of ISO-Space labels as input for a pipeline system. A second system of note was created by researchers from the University of the Basque Country (UPV-EHU) also employs an SVM classifier, in addition to a pipeline architecture that employs ClearNLP, OpenNLP, WordNet, PropBank, and the Predicate Matrix (Salaberri 2015). This system’s successful use of WordNet inspired the methodology proposed to extract ‘measure’ labels that is reported in the following section.

**Methodology**

A rule-based system was constructed to identify the most common constructions of “measures” according to the ISO-Space scheme. All documents were tokenized using NLTK’s built-in tokenizer function, stripped of punctuation, and converted to lowercase. For each document, sentences were parsed using the NLTK “pos\_tag” function to obtain estimates of part-of-speech for each token.

The rule-based system relies first upon finding candidate words of interest. In order to identify the majority of “measure” tags included in the training data, the program searches for all instances of the ‘measure.n.02’ synset from WordNet in the target document. A function recursively discovers all hyponyms of the ‘measure.n.02’ from which possible candidate ‘measure’ strings may be identified. The hyponyms were then pruned to exclude all synsets that likely connoted non-spatial measures, including any measures of time or money. This was achieved by identifying keywords pertaining to time and currency in the glosses of the hyponym synsets and removing these senses from the candidate list. Those words that excluded a synset from inclusion in the candidate set included the following: {'volume', 'contained', 'contain', 'time', 'hold', 'period'}. Additionally, the glosses of all candidates were pruned such that one of the following words was present in the gloss: {'distance', 'space', 'quantity', 'length', 'width'}. By pruning the synsets in this manner, potential measurement words that were not related to distance measures were excluded. Lemmas for each of the remaining synsets were then added to a list of possible manifestations of the “measure.n.02” synset pertaining to space and distance. While nearly all of these candidates were left in the final list, the candidate lemma “in” (an abbreviation of the word “inch”) was removed, as it coincides with the common preposition that does not denote a “measure”. The initial candidate list of lemmas was then expanded to include pluralized forms of the respective words. This was achieved with the “inflect” package in Python. Let this list be called INFLIST.

After this initial candidate list of words was identified, each document was scanned for the presence of any word in the candidate list. Then, a series of rules are applied to determine the full span of the measure. Two additional lists were compiled to identify words in spans containing measures, including a basic list (‘adj’) of relative distance words (e.g. ‘near’, ‘close’, ‘far’, ‘nearby’) and a list of adjectives (‘degreeadjs’) describing the degree of a potential distance (e.g. ‘many’, ‘several’, ‘exactly’, ‘about’, ‘approximately’, etc.). The basic list of rules compiled to add to the list of candidate measure spans is printed below (does not include preprocessing rules):

Let WORD be the current word:

1. IF WORD is in ‘adj’ and prevT is a verb, add WORD.

ELIF WORD is in {"near", "nearby", "close"}, add WORD.

1. IF WORD in INFLIST and prevW is a number, add (prevW, WORD).
2. IF WORD in INFLIST and prevW is a number and prev2W is in ‘degreeadjs’, add (prev2W, prevW, WORD).
3. IF WORD in INFLIST and prevT is ‘DT’ and prev3W and prev2W form a comparative phrase, add (prev3W, prev2W, prevW, WORD).

These rules were constructed in an iterative process in order to maximize the recall score of the predictions against the training data gold standard measure labels. The F1 score and precision were viewed as desirable, but less so than recall, as the primary task is to identify all possible ‘measure’ labels from the data. Development data was not separated out from the training data due to the low number of instances of the measure label in the training data to begin with.

**Results**

The four rules listed above were enough to achieve 77% recall and 87.5% recall on the test data. Full results of the system on both the train and test data are reported below:

Training Data:

finalrecall 0.7705882352941177

finallprec 0.6359223300970874

finalF 0.696808510638298

Test Data:

finalrecall 0.875

finallprec 0.4794520547945205

finalF 0.6194690265486725

Further analysis of the output of the system reveals that many of the discrepancies between the predicted span set and the gold standard ‘measure’ spans were resultant from inconsistencies in annotation of the gold standard data. For example, the simple system for extracting ‘measure’ spans picks up the span ‘some 20 km’, indicating an *approximate* distance of 20 kilometers, whereas the gold standard span includes only ’20 km’ in the span. This is inconsistent with decisions elsewhere in the training data, in which spans of the form [*quantifier*, *number*, *distance\_measure\_word*] are all included in the span of the gold standard label for a ‘measure’ tag. The proposed ‘measure’ extraction system would likely achieve higher results if annotation conventions were consistent across all data. While inter-annotator agreement for ‘measure’ tags was not reported, the authors of the SpaceEval task note low Fleiss scores for ‘OLink’ and ‘QLink’ spatial relation tags (<40%). This may suggest that the annotation task was not consistently carried out.

**Conclusion and Future Work**

This work suggests that very few rules can determine the majority of “measure” tags as encoded in the ISO-Space annotation scheme for the type of training data provided (trip review narratives). However, the proposed system did not seriously account for alternate spellings, misspellings, and abbreviations (e.g. ‘approx..’), conjunctive phrases (e.g. ‘between 6,000 and 8,000 ft’), non-quantitative modifiers of spatial words (e.g. ’30 flat miles’), elisions of implied spatial words, and sub-trees that contain many modifiers. A more flexible rule-based approach would be one that identified the sub-trees in which spatial candidate words are found, and to prune the sub-tree. This would require an inversion of the rule-creation approach, such that pruning rules are preferred over affirmative rules that observe previous n-grams. While the above oversights were not implemented due to time constraints, it is straightforward to implement rules to cover these cases to attain higher recall in a revised system. However, certain phrases and words in context marked as ‘measures’ that were not identified by the system are more difficult to write concrete rules for (e.g. ‘within sufficient accuracy’, ‘just’ (as in ‘just to the north’)).

A learning approach based on context that uses trees as a feature, as proposed by D’Souza, would likely obtain higher scores for this task if a larger training set were available. Future work in this area should include special attention to tree structures to maximize the likelihood that the entire relevant span is tagged, including sub-trees that involve long modifying phrases and conjunctive clauses. Output from this model may be able to serve as a feature for future learning models designed to predict “MLink” relations, as well as the attributes of other relations in ISO-Space.

**Sources**

D'Souza, J., Ng, Vincent, Gupta, Gopal, Liu, Yang, & Tamil, Lakshman. (2015). *Extracting Time and Space Relations from Natural Language Text,* ProQuest Dissertations and Theses.

D'Souza, J., & Ng, V. (2015). UTD: Ensemble-based spatial relation extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 862-869).

D'Souza, J., & Ng, V. (2015). Sieve-based spatial relation extraction with expanding parse trees. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 758-768).

Kordjamshidi, P., Rahgooy, T., & Manzoor, U. (2017). Spatial Language Understanding with Multimodal Graphs using Declarative Learning based Programming. In *Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing* (pp. 33-43).

Martins, B., López-Pellicer, F. J., & Ahlers, D. (2015, November). Expanding the utility of geospatial knowledge bases by linking concepts to WikiText and to polygonal boundaries. In *Proceedings of the 9th Workshop on Geographic Information Retrieval* (p. 1). ACM.

Mazalov, A., Martins, B., & Matos, D. (2015, November). Spatial role labeling with convolutional neural networks. In *Proceedings of the 9th Workshop on Geographic Information Retrieval* (p. 12). ACM.

Ramisa Ayats, A., Wang, J., Lu, Y., Dellandrea, E., Moreno-Noguer, F., & Gaizauskas, R. (2016). Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 EMNLP Conference on Empirical Methods in Natural Language Processing* (pp. 214-220).

Salaberri, H., Arregi, O., & Zapirain, B. (2015). IXAGroupEHUSpaceEval:(X-Space) A WordNet-based approach towards the automatic recognition of spatial information following the ISO-Space annotation scheme. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 856-861).

Pustejovsky, J., Moszkowicz, J. L., & Verhagen, M. (2011). Using ISO-Space for annotating spatial information. In *Proceedings of the International Conference on Spatial Information Theory*.

Pustejovsky, J., Moszkowitz,, J. Verhagen, M. A linguistically grounded annotation language for spatial information. (2012). http://www.atala.org/sites/default/files/3-Pustejovsky-TAL53-2.pdf

Pustejovsky, J., Kordjamshidi, P., Moens, M. F., Levine, A., Dworman, S., & Yocum, Z. (2015). SemEval-2015 Task 8: SpaceEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)* (pp. 884-894).