Semantic role labeling

Yukang Yang

1. Introduction

In this project, I mainly deal with semantic roles covered in The Proposition Bank (PropBank). The semantic roles covered by PropBank include:

- **Numbered arguments (A0–A5, AA)**: Their semantics depends on the verb and the verb usage in a sentence. Commonly, A0 stands for the agent and A1 corresponds to the patient or theme of the proposition.
- **Adjuncts (AM-)**:
  AM-ADV : general-purpose
  AM-CAU : cause
  AM-DIR : direction
  AM-DIS : discourse marker
  AM-EXT : extent
  AM-LOC : location
  AM-MNR : manner
  AM-MOD : modal verb
  AM-NEG : negation marker
  AM-PNC : purpose
  AM-PRD : predication
  AM-REC : reciprocal
  AM-TMP : temporal
- **References (R-)**: The role of a reference is the same as the role of the referenced argument.
- **Verbs (V)**

2. Dataset

I use dataset from CoNLL 2005 shared task to train and evaluate. You can download data from its website: http://www.lsi.upc.edu/~srlconll/soft.html. Besides data in the website, you also need to download the word package from http://www.lsi.upc.edu/~srlconll/conll05st-release-words.tar.gz, which is a part of dataset but is not shown in the website.

3. SRL system architecture

The system consists of four stages: pruning, argument identification, argument classification, and inference.

3.1 Pruning

Pruning stage decreases number of candidates for next stage by eliminating impossible candidates. This stage keeps all possible candidates and therefore leads to small false negative and high recall. The system exploits the heuristic rules introduced by Xue and Palmer (2004) to filter out simple constituents that are very unlikely to be arguments.

Step1: Designate the predicate as the current node and collect its sisters (constituents attached at the same level as the predicate)
Step 2: Reset the current node to its parent and repeat Step 1 till it reaches the top level node.

I get about 91% recall with this rule. If the syntactic parsing is 100% accurate, this rule should get 100% recall but the truth is about 10% syntactic parsing is wrong. The training set was parsed by a older version of Charniak parser. To improve syntactic parsing quality, I used the latest version of Charniak parser to reparse the dataset. This improves recall by 1%. I trace down errors and add following rules:

Rule 1: Add children of preposition phrase and clause
Rule 2: Concatenate consecutive noun phrase and preposition phrase

After that, I got 96% recall.

I also remove candidates that overlap with the predicate. This is added as a linguistic constraints.

3.2 Identification

The identification stage utilizes binary classification to identify whether a candidate is an argument or not. I chose stochastic gradient descent (SGD) classifier with 40 iterations.

3.3 Classification

This stage assigns labels to the argument candidates identified. A multi-class classifier is trained to predict the types of the argument candidates. I still used SGD classifier with 40 iterations.

3.4 features

Features are needed for classifiers. I collect 54 features from previous research (Pradhan et al., 2004; Xue and Palmer, 2004; Haghighi et al., 2005; Pradhan et al., 2005; Punyakanok et al., 2007). I will compare my features with Punyakanok's by setting the common part as baseline.

3.4.1 Baseline features

Following features are the common part between my system and Punyakanok's.

- **Predicate**: lemma of the predicate.
- **Candidate phrase type**
- **Parse tree path**: The path in parse tree from the constituent being classified to the predicate. This feature is important for identification but is not effective for classification.
- **Position**: Position The relative position of the constituent being classified with regard to the predicate (before or after).
- **Voice**: Whether the predicate is active or passive.
- **Candidate head word and head word POS tag**
- **Sub-categorization**: The phrase structure rule expanding the parent of the predicate.
- **Context words and POS tags of the context words**: the feature includes the two words before and after the argument, and their POS tags

Baseline F-measure:

R-A0: 0.891402714932	A3: 0.493927125506	AM-LOC: 0.522658610272
R-A1: 0.747252747253	A2: 0.598441558442	AM-MOD: 0.991507430998
R-A2: 0.571428571429	A4: 0.690058479532	AM-CAU: 0.578947368421
AM-MNR: 0.56420233463	AM-DIR: 0.333333333333	C-A1: 0.674816625917
AM-EXT: 0.588235294118	AM-DIS: 0.727848101266	AM-ADV: 0.589595375723
A1: 0.786339421613	AM-TMP: 0.694570135747	Identifier: 0.945371460223
A0: 0.869979996923	AM-PNC: 0.419263456091	AM-NEG: 0.991689750693

3.4.2 Other features that I used in this project

These features in my project are different from Punyakanok's. ***For each feature, I give a table that shows the change in F-measure after adding that feature to baseline.*** In the table, I only show changes that are more than 0.009.

- **First and last word and their POS tags in constituent**

R-A0 : -0.009682284824	AM-DIR : 0.107843137255
R-A1 : -0.014858381056	AM-DIS : 0.104978646454
R-A2 : 0.043956043956	AM-TMP : 0.102924645046
AM-MNR : 0.026706756279	AM-PNC : 0.017711333825
AM-EXT : 0.024009603841	AM-LOC : 0.110368912664
A1 : 0.011176106337	AM-CAU : 0.08142999007
A2 : 0.038400546821	C-A1 : -0.009700346847
A4 : 0.078234203395	AM-ADV : 0.059365363307

- **Constituent relative features**: Nine features representing the phrase type, head word and head word part of speech of the parent, and left and right siblings of the constituent.

R-A0 : -0.05772715025	A2 : 0.037968436459
R-A1 : -0.051007013465	AM-DIR : 0.066666666667
R-A2 : -0.071428571429	AM-PNC : 0.023214420015
AM-MNR : 0.011140131123	AM-LOC : 0.01580292819
AM-EXT : 0.063938618925	AM-CAU : 0.03216374269
A1 : 0.043860401204	C-A1 : 0.019358131365
A3 : 0.044225484936	AM-ADV : 0.022649522236

- **Head of PP parent**: If the parent of the current constituent is a PP, then the head of this PP, the preposition is added as a feature.

  No change that is more than 0.009

- **Partial path**: Path from the constituent to the lowest common ancestor of the predicate and the constituent. This is a variance of path feature.

R-A0 : -0.089142827926	AM-DIR : 0.04347826087
R-A1 : -0.103184950643	AM-CAU : -0.027766266059
R-A2 : -0.15037593985	AM-ADV : -0.019125577065

- **PP NP Head-Word/Head-POS**: For a PP, retrieve the head word and head POS of head NP in it.

R-A2 : 0.043956043956	A2 : 0.024300205639
AM-MNR : 0.025835979546	A4 : 0.048036758563
AM-EXT : 0.036764705882	AM-DIR : 0.04347826087

AM-DIS : 0.067332621626
AM-TMP : 0.099012751954
AM-PNC : 0.010969102049
AM-LOC : 0.124308730319
AM-CAU : 0.122807017544

- **Length of the Path feature**

R-A2 : -0.071428571429
R-AM-TMP : 0.642857142857
AM-EXT : -0.022197558269
A3 : 0.017700781471
AM-DIR : -0.009803921568
AM-DIS : 0.019304518324
AM-CAU : -0.026101839966
AM-ADV : 0.009786704607

- **Predicate lemma & head word**

R-A1 : 0.009110889111
R-A2 : 0.043956043956
AM-MNR : 0.012647760247
AM-EXT : 0.0935828877
A1 : 0.012591774112
A3 : 0.125475859569
A2 : 0.05001570003
A4 : 0.065755473956
AM-DIR : 0.032863849766
AM-TMP : 0.018225031743
AM-PNC : 0.049486543909
AM-LOC : 0.038515941115
AM-ADV : 0.014688727997

- **Predicate lemma & phrase type**

R-A0 : 0.012706874109
R-A2 : -0.038095238096
AM-MNR : 0.028539600854
AM-EXT : 0.064825930372
A1 : 0.021593772541
A3 : 0.064043888987
A2 : 0.049148803004
AM-DIR : -0.047619047619
AM-DIS : -0.012658227848
AM-TMP : 0.010393612162
A4 : 0.071305156832
AM-PNC : 0.014754139217
AM-LOC : 0.036120015682
AM-CAU : 0.04241185488
C-A1 : 0.016598687308
AM-ADV : -0.009548972011

- **Voice & position**

R-A1 : -0.01079065339
R-A2 : 0.043956043956
AM-MNR : 0.012795716052
AM-EXT : 0.011764705882
AM-DIR : 0.04347826087
AM-CAU : -0.022425629291

- **Predicate Lemma & PP Parent Head-Word**

R-A2 : -0.038095238096
AM-EXT : -0.011312217195

AM-CAU : -0.02722323049

3.4.3 Conclusion

As we can see from these data, following features are effective for improving classification output:

- **Predicate lemma & phrase type** (By &, I mean concatenating two features and therefore form a new feature)
- **Predicate lemma & head word**
- **Constituent relative features** (Nine features representing the phrase type, head word and head word part of speech of the parent, and left and right siblings of the constituent.)
- **PP NP Head-Word/Head-POS** (For a PP, retrieve the head word and head POS of head NP in it. This feature is different from head word of PP because head word of PP is its dominant preposition.)
- **First and last word and their POS tags in constituent**

Following features increase F-measure by over 10% which means they are very effective:

- **First and last word and their POS tags in constituent**
- **Predicate lemma & head word** (By &, I mean concatenating two features and therefore form a new feature)
- **PP NP Head-Word/Head-POS** (explained before)

Following features should be removed:

- **Head of PP parent** (makes no change at all)
- **Partial path** (makes too many negative changes)

3.4.4 Other features that I haven't used but might be helpful

- **Predicate Lemma & Path**
- **Is predicate plural & is noun phrase plural**: these two features together might help improve labeling of A0.
- **Candidate length**: the number of words in the candidate
- **Syntactic frame**: This feature represents the syntactic structure around candidate. For example, “I like dogs.” The syntactic structure is “np v np”. The syntactic frame for “dog” is “np v NP”. In this way, we can know the syntactic structure and the position of candidate in this structure.
- **Verb class**: verb class in PropBank
- **Temporal Keywords**: Is the head of the node a temporal word e.g. ‘February’ or ‘afternoon’
- **Missing subject**: Is the predicate missing a subject in the “standard” location
- **Whether the constituent is in the same clause as the predicate**

3.5 Inference

This stage applies linguistic constraints to output from classification stage.
Some constraints:

- Candidates can't overlap with each other.
- Usually A0-A5 occur only once in a sentence

4 Evaluation

4.1 Pruning evaluation

Recall: 0.960959277242
Precision: 0.299537548324
F-measure: 0.456714178261

4.2 Identification evaluation

Recall: 0.972665836071
Precision: 0.949298058145
F-measure: 0.960839891098

4.3 Classification evaluation

Type	Precision	Recall	F-measure
A0	97.57%	83.56%	90.02%
A1	95.68%	80.53%	87.45%
A2	82.10%	69.46%	75.25%
A3	69.63%	69.12%	69.37%
A4	88.60%	72.16%	79.55%
R-A0	98.00%	85.22%	91.16%
R-A1	84.67%	71.17%	77.33%
AM-ADV	78.60%	59.25%	67.57%
AM-CAU	77.36%	67.21%	71.93%
AM-DIR	60.00%	47.72%	53.16%
AM-DIS	88.49%	77.16%	82.44%
AM-EXT	53.57%	83.33%	65.22%
AM-LOC	87.40%	61.14%	71.95%
AM-MOD	100%	98.06%	99.02%
AM-MNR	65.36%	62.46%	63.87%
AM-NEG	99.53%	98.15%	98.53%
AM-PNC	67.01%	43.96%	53.57%
AM-TMP	93.13%	76.17%	83.79%
C-A1	69.24%	73.95%	70.98%

4.4 Comparison

Type	My F-measure	Punyakanok's
A0	90.02%	88.05%
A1	87.45%	79.91%
A2	75.25%	68.16%
A3	69.37%	64.31%
A4	79.55%	77.25%

AM-ADV	67.57%	59.73%
AM-CAU	71.93%	53.97%
AM-DIR	53.16%	46.48%
AM-DIS	82.44%	77.95%
AM-EXT	65.22%	55.56%
AM-LOC	71.95%	60.33%
AM-MNR	63.87%	59.22%
AM-NEG	98.83%	97.61%
AM-PNC	53.37%	45.41%
AM-MOD	99.02%	97.40%
R-A0	91.16%	87.67%
R-A1	77.33%	73.01%
R-A2	62.5%	52.17%
R-AM-LOC	91.67%	68.57%
R-AM-TMP	75.86%	68.69%

The comparison shows that my program gets better result than the one by Punyakanok. I think other features that I collect help me get a better result as you can see from the feature section.

References

Michael Collins. 1999. Head-driven Statistical Models for Natural Language Parsing. Ph.D. thesis, University of Pennsylvania.

Aria Haghighi, Kristina Toutanova and Christopher Manning, A Joint Model for Semantic Role Labeling.

Sameer Pradhan, Kadri Hacioglu, Wayne Ward, James H. Martin and Daniel Jurafsky, Semantic Role Chunking Combining Complementary Syntactic Views.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. 2004. Shallow semantic parsing using support vector machines. In Proceedings of the Human Language Technology Conference/North American chapter of the Association of Computational Linguistics (HLT/NAACL), Boston, MA.

V.Punyakanok and D. Roth and W. Yih, The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. Computational Linguistics (2008)