Relation extraction

Bill MacCartney
CS224u
Stanford University

Directions to explore

Overview

- ~~The task of relation extraction~~
- ~~Data resources~~
- ~~Problem formulation~~
- ~~Evaluation~~
- ~~Simple baselines~~
- Directions to explore

Directions to explore

- Examining the trained models
- Discovering new relation instances
- Enhancing the model

Directions to explore

Examining the trained models

```
rel_ext.examine_model_weights(train_result)
```

```
Highest and lowest feature weights for relation author:

     3.055 author
     3.032 books
     2.342 by
     ..... .....
    -2.002 directed
    -2.019 or
    -2.211 poetry
```

```
Highest and lowest feature weights for relation adjoins:

     2.511 Córdoba
     2.467 Taluks
     2.434 Valais
     ..... .....
    -1.143 for
    -1.186 Egypt
    -1.277 America
```

```
Highest and lowest feature weights for relation
film_performance:

     4.004 starring
     3.731 alongside
     3.199 opposite
     ..... .....
    -1.702 then
    -1.840 She
    -1.889 Genghis
```

```
Highest and lowest feature weights for relation has_spouse:

     5.319 wife
     4.652 married
     4.617 husband
     ..... .....
    -1.528 between
    -1.559 MTV
    -1.599 Terri
```

Directions to explore

Discovering new relation instances

```
rel_ext.find_new_relation_instances(
    dataset,
    featurizers=[simple_bag_of_words_featurizer])
```

```
Highest probability examples for relation adjoins:

     1.000 KBTriple(rel='adjoins', sbj='Canada', obj='Vancouver')
     1.000 KBTriple(rel='adjoins', sbj='Vancouver', obj='Canada')
     1.000 KBTriple(rel='adjoins', sbj='Australia', obj='Sydney')
     1.000 KBTriple(rel='adjoins', sbj='Sydney', obj='Australia')
     1.000 KBTriple(rel='adjoins', sbj='Mexico', obj='Atlantic_Ocean')
     1.000 KBTriple(rel='adjoins', sbj='Atlantic_Ocean', obj='Mexico')
     1.000 KBTriple(rel='adjoins', sbj='Dubai', obj='United_Arab_Emirates')
     1.000 KBTriple(rel='adjoins', sbj='United_Arab_Emirates', obj='Dubai')
     1.000 KBTriple(rel='adjoins', sbj='Sydney', obj='New_South_Wales')
     1.000 KBTriple(rel='adjoins', sbj='New_South_Wales', obj='Sydney')
```

Directions to explore

Discovering new relation instances

```
rel_ext.find_new_relation_instances(
    dataset,
    featurizers=[simple_bag_of_words_featurizer])
```

```
Highest probability examples for relation author:

     1.000 KBTriple(rel='author', sbj='Oliver_Twist', obj='Charles_Dickens')
     1.000 KBTriple(rel='author', sbj='Jane_Austen', obj='Pride_and_Prejudice')
     1.000 KBTriple(rel='author', sbj='Iliad', obj='Homer')
     1.000 KBTriple(rel='author', sbj='Divine_Comedy', obj='Dante_Alighieri')
     1.000 KBTriple(rel='author', sbj='Pride_and_Prejudice', obj='Jane_Austen')
     1.000 KBTriple(rel='author', sbj="Euclid's_Elements", obj='Euclid')
     1.000 KBTriple(rel='author', sbj='Aldous_Huxley', obj='The_Doors_of_Perception')
     1.000 KBTriple(rel='author', sbj="Uncle_Tom's_Cabin", obj='Harriet_Beecher_Stowe')
     1.000 KBTriple(rel='author', sbj='Ray_Bradbury', obj='Fahrenheit_451')
     1.000 KBTriple(rel='author', sbj='A_Christmas_Carol', obj='Charles_Dickens')
```

Directions to explore

Discovering new relation instances

```
rel_ext.find_new_relation_instances(
    dataset,
    featurizers=[simple_bag_of_words_featurizer])
```

```
Highest probability examples for relation capital:

     1.000 KBTriple(rel='capital', sbj='Delhi', obj='India')
     1.000 KBTriple(rel='capital', sbj='Bangladesh', obj='Dhaka')
     1.000 KBTriple(rel='capital', sbj='India', obj='Delhi')
     1.000 KBTriple(rel='capital', sbj='Lucknow', obj='Uttar_Pradesh')
     1.000 KBTriple(rel='capital', sbj='Chengdu', obj='Sichuan')
     1.000 KBTriple(rel='capital', sbj='Dhaka', obj='Bangladesh')
     1.000 KBTriple(rel='capital', sbj='Uttar_Pradesh', obj='Lucknow')
     1.000 KBTriple(rel='capital', sbj='Sichuan', obj='Chengdu')
     1.000 KBTriple(rel='capital', sbj='Bandung', obj='West_Java')
     1.000 KBTriple(rel='capital', sbj='West_Java', obj='Bandung')
```

Directions to explore

Discovering new relation instances

```
rel_ext.find_new_relation_instances(
    dataset,
    featurizers=[simple_bag_of_words_featurizer])
```

```
Highest probability examples for relation worked_at:

     1.000 KBTriple(rel='worked_at', sbj='William_C._Durant', obj='Louis_Chevrolet')
     1.000 KBTriple(rel='worked_at', sbj='Louis_Chevrolet', obj='William_C._Durant')
     1.000 KBTriple(rel='worked_at', sbj='Iliad', obj='Homer')
     1.000 KBTriple(rel='worked_at', sbj='Homer', obj='Iliad')
     1.000 KBTriple(rel='worked_at', sbj='Marvel_Comics', obj='Stan_Lee')
     1.000 KBTriple(rel='worked_at', sbj='Stan_Lee', obj='Marvel_Comics')
     1.000 KBTriple(rel='worked_at', sbj='Mongol_Empire', obj='Genghis_Khan')
     1.000 KBTriple(rel='worked_at', sbj='Genghis_Khan', obj='Mongol_Empire')
     1.000 KBTriple(rel='worked_at', sbj='Comic_book', obj='Marvel_Comics')
     1.000 KBTriple(rel='worked_at', sbj='Marvel_Comics', obj='Comic_book')
```

Directions to explore

Error analysis

```
exs = dataset.corpus.get_examples_for_entities( 'Louis_Chevrolet' , 'William_C._Durant' )
for ex in exs:
    print(' | '.join((ex.left[ -10:], ex.mention_1, ex.middle, ex.mention_2, ex.right[:  10])))
```

```
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
Founded by | Louis Chevrolet | and ousted GM founder | William C. Durant | on Novembe
```

```
model = train_result[ 'models' ]['worked_at' ]
vectorizer = train_result[ 'vectorizer' ]
print(model.coef_[0][vectorizer.vocabulary_[ 'founder' ]])
```

```
2.0528435038145383
```

Directions to explore

Error analysis

```
print(len(dataset.corpus.get_examples_for_entities( 'Homer', 'Iliad')))
```

```
118
```

```
mids = defaultdict(int)
for ex in dataset.corpus.get_examples_for_entities( 'Homer', 'Iliad'):
    mids[ex.middle]  += 1
for cnt, mid in sorted([(cnt, mid)  for mid, cnt in mids.items()], reverse =True)[:5]:
    print('{:10d} {:s}'.format(cnt, mid))
```

```
        51 's
        13 ' s
         4 , and in particular the
         4 ,
         3 in the
```

```
model = train_result['models']['worked_at']
vectorizer = train_result['vectorizer']
print(model.coef_[0][vectorizer.vocabulary_[ "'s"]])
```

```
0.5801433006163413
```

Directions to explore

Enhancing the model: feature representations

- Word embeddings
- Directional bag-of-words
- N-grams
- POS tags
- WordNet synsets
- Syntactic features
- Features based on entity mentions
- Features based on `left` and `right`

Directions to explore

Enhancing the model: model types

- Support vector machines (SVMs)
- Feed-forward neural networks
- LSTMs
- Transformers