PAPER USING WEB AS CORPUS

1. Copus:
   1. proses ini mengambil contoh-contoh yang belum dilabeli dari web secara otomatis
   2. bangun query dengan cara
   3. kombinasikan most significant words untuk setiap sense dari kata yang polisemi
   4. gunakan query tersebut cari di web untuk tambahan contoh training yang berhubungan dengan given sense tersebut
   5. Membangun query:
      1. tentukan set of relevant words dari setiap sense dalam training korpus
      2. kita menganggap word i relevant pada sense S jika:
         1. frequensi wi dalam S lebih besar dari rata-rata kemunculan semua kata (yang lebih dari sekali) dalam sense itu.
         2. Information gain wi pada training set adalah positif. Untuk memilih kata yang mengurangi ketidakpastian dari value sebuah sense dari banyak contoh-contoh
      3. Dari kumpulan relevant words per sense tersebut, buat querynya.
      4. Buat query dengan tiga kata
      5. Query didapat dari kominasi relevant words per sense sepanjang tiga buah kata
   6. Web searching:
      1. menggunakan query untuk mengambil unlabeled text dari web
      2. Setiap hasil download dapat dikatakan bahwa berkaitan dengan suatu sense
   7. Semi Supervised learning:
      1. meningkatkan akurasi dari klasifikasi dengan membesarkan set training asli yang sedikit, dengan contoh unlabeled text hasil download dari Web.
      2. Pada awalnya, classifier ditraining dengan set data original (labeled) yang sedikit.
      3. Classifier ini kemudian mengklasifikasikan unlabeled text data.
      4. Contoh dengan tingkat kepercayaan tinggi disambungkan dengan label hasil prediksinya, ditambahkan ke training set.
      5. Seleksi dari hasil dengan tingkat kepercayaan tinggi tidak hanya didapat dari tingkat kepercayaan classifier tapi juga dari Web based labeling.
   8. Evaluasi:
      1. diambil dari subset of contoh leksikal task dari SemEval forum.
      2. Gunakan 9 kata benda yang punya training data untuk setiap sensenya.
      3. Walaupun ada kata tertentu yang punya jumlah training tidak seimbang antar sensenya.

One Million Sense-Tagged Instancesfor Word Sense Disambiguation and Induction

Training Data:

1. Extract dan tag sense data dari paralel korpus dengan cara semi otomatis
2. Menggunakan MultiUN korpus.
3. Menggunakan Chinese-English dari korpus tersebut
4. Proses ekstraksi:
   1. Tokenisasi masing-masing korpus (inggris dan chinese)
   2. GIZA++ digunakan untuk word alignment antara kata-kata dalam chinese dan english
   3. POS tagging dan lematisasi dari kalimat bahasa inggris
   4. Annotasi:
      1. Assign sense dari wordnet ke kata w dalam bahasa inggris
      2. Gunakan wc yang merupakan translasi dalam bahasa chinese dari alignment tadi
      3. Untuk setiap sense i dari w (english), buat list dari hasil translate tiap sense i dalam bahasa china (manual)
      4. Jika wc sama dengan salah satu list dari sense i dalam bahasa china tersebut, maka w ditag dengan sense i

Evaluasi:

Using IMS WSD system

* SenseEval-2
* SenseEval-3
* Fine grained Sem-Eval-2007
* Coarse grained Sem-Eval-2007
* SemEval-2013 WSI task, 50 lemma, 20 noun, 30 verb, 10 adjektif, anotasi dengan wordnet 3.1

Towards Automatic Acquisition of a Fully Sense Tagged Corpus for Persian

Pembuatan Sense Tagged Corpus:

1. Menggunakan paralel corpus english-persian
2. Setelah itu, gunakan direct knowledge based algorithm untuk tagging sense ke kata-kata sisanya.
3. Kedua bahasa dilakukan POS Tagging dan lemmetisasi
4. Menggunakan pendekatan cross lingual untuk tag sense kata dalam Persian text.
5. Tag yang dalam bahasa inggris menggunakan Extended Lesk Algorithm, jika ada tag-tag yang salah, maka akan dibetulkan terlebih dahulu
6. Tag semua kata ambigu di input berbahasa inggris.
7. Transfer synset dari english ke persian menggunakan lingual relations by FarstNet dan mencocokkan synset wordnet english yang diassign ke suatu kata ke corresponding sysnet di FarsNet.
8. Tidak menggunakan word alignment karena untuk akurasi yang kurang baik bisa berakibat fatal
9. Untuk setiap synset dari butir 7 yang mengandung kata-kata synonim dalam Persian, assign semua kata tersebut dengan sense yang sama dengan english label translasinya.
10. Untuk menambah synset dari sisi Farsi, untuk setiap english tagged word, cari translasinya, tambahkan ke Farsi synset.
11. Tambahan tag dengan ExtLesk pada korpus diambil yang memenuhi confidence threshold minimal

Evaluasi:

1. tagged korpus dievaluasi dengan 480 words random dari berbagai bidang yang mempunyai rata-rata jumlah sense 2.17
2. 7 expert linguistik dilibatkan dalam evaluasi
3. Pada proses tagging english-side tag-tag yang salah dibenarkan secara manual
4. Hasilnya memuaskan