############################Quotes-quotes penting###############################

“Through word sense disambiguation ex- periments performed on the Wikipedia-based sense tagged corpus generated for a subset of the SENSE- VAL ambiguous words, we show that the Wikipedia annotations are reliable, and the quality of a sense tagging classifier built on this data set exceeds by a large margin the accuracy of an informed baseline that selects the most frequent word sense by default. “ – Rada

“Wikipedia is a free online encyclopedia, represent- ing the outcome of a continuous collaborative effort of a large number of volunteer contributors.” – Rada

“One of the implications of the large number of contributors editing the Wikipedia articles is the occasional lack of consistency with respect to the unique identifier used for a certain entity. For in- stance, the concept of *circuit (electric)* is also re- ferred to as *electronic circuit*, *integrated circuit*, *electric circuit*, and others. This has led to the so- called *redirect pages*, which consist of a redirection hyperlink from an alternative name (e.g. *integrated circuit*) to the article actually containing the descrip- tion of the entity (e.g. *circuit (electric)*).” – Rada

Intinya karena kontributornya banyak, ada inkonsisten dalam penamaan entity yang terkadang beda nama namun merujuk pada hal yang sama. Karena itu ada yang namanya redirect page.

Ada juga halaman ambigu, diperuntukan untuk entitas yang ambigu dan mempunyai banyak makna tergantung dari konteks. Sebagai uniq identifier terdapat pendanda (disambiguation) untuk merefer link ke halaman ambigu tersebut.

“Alternatively, techniques have been proposed for discovering senses of words automatically from unannotated text. This task of unsupervised word sense induction (WSI) can be conceptualized as a clustering problem. To correctly identify all senses of polysemous words encountered in a corpus, words can be clustered according to their meanings and allowing multiple membership. “ – Michael

“Word sense disambiguation (WSD) is the ability to identify the meaning of words in context in a compu- tational manner. WSD is considered an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence. “ – Roberto

“Human language is ambiguous, so that many words can be interpreted in multiple ways depending on the context in which they occur. “ – Roberto

“Second, WSD heavily relies on knowledge. In fact, the skeletal procedure of any WSD system can be summarized as follows: given a set of words (e.g., a sentence or a bag of words), a technique is applied which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context. Knowl- edge sources can vary considerably from corpora (i.e., collections) of texts, either unla- beled or annotated with word senses, to more structured resources, such as machine- readable dictionaries, semantic networks, etc. “ – Roberto

Karena kebutuhan resource tersebut lah maka kita akan generate korpus dari Wikipedia, dikarenakan resource misalnya wordnet untuk Bahasa Indonesia yang masih terbatas.

“Unfortunately, the manual creation of knowledge resources is an expensive and time- consuming effort [Ng 1997], which must be repeated every time the disambiguation scenario changes (e.g., in the presence of new domains, different languages, and even sense inventories). This is a fundamental problem which pervades the field of WSD, and is called the *knowledge acquisition bottleneck* [Gale et al. 1992b]. “ – Roberto

“The potential of WSD is also clear when we deal with the problem of machine transla- tion: for instance, the Italian word *penna* can be translated in English as *feather*, *pen*, or *author* depending upon the context. “ – Roberto

Terdapat 2 varian WSD Task: [Roberto]

* Lexical sample (targeted WSD): scope dari kata-kata dibatasi
* All words: semua text bisa didisambiguasi namun demikian ini susah karena data training dan resource yang dibutuhkan sangat banyak dan kemungkinan sulit untuk mengcover semua kata-katanya.

“A *word sense* is a commonly accepted meaning of a word. “ – Roberto

Meaning dari suatu kata juga sebenarnya sulit untuk didiscretekan jika kita ingin melihat WSD task sebagai task klasifikasi dari kata ke sense yang tepat. Karena interpretasi ‘meaning’ dari kata merupakan hal yang kompleks dan tergantung dari sudut pandang manusia yang menilainya.

“The goal of a WSD is to computationally assign the correct sense of a word (i.e. meaning) in context (phrase, sentence, paragraph, text) from a predefined sense inventory, when the word has multiple meanings. “ – Nasiruddin

Terdapat berbagai macam resource yang ada untuk membantu proses WSD, dari wordnet, thesaurus, korpus seperti brown corpus, semcor, dll.

“A concept in WordNet is represented as a synonym set, or *synset*, i.e. the set of words which share a common meaning. pop , soda , soda pop , soda water , tonic. where each word’s subscripts and superscripts in- dicate their parts of speech (e.g. stands for noun) and sense number1, respectively. For each synset, WordNet provides a textual definition, or gloss. For example, the gloss of the above synset is: “a sweet drink containing carbonated water and fla- voring”. “ – Simone Paolo

“ORED WATER, LEMON- ADE, and many others. The title of a Wikipage (e.g. SODA (SOFT DRINK)) is composed of the lemma of the concept defined (e.g. soda) plus an optional label in parentheses which specifies its meaning in case the lemma is ambiguous (e.g. SOFT DRINK vs. SODIUM CARBONATE).

“ – Simone Paolo

################## Using Wikipedia for Automatic Word Sense Disambiguation##########

Cara Pada Paper [Rada] (Using Wikipedia for Automatic Word Sense Disambiguation):

Given kata-kata yang ambigu, cara bikin sense-tagged korpusnya:

Ambil semua paragraph di Wikipedia yang mengandung kemunculan kata-kata ambigu tersebut sebagai bagian dari link atau piped link. Ambil one paragraph per line (sekitar 80 kata). Hindari named entity dengan hanya mengambil yang huruf kecil saja. Ambil label dari kata ambigu tersebut dengan mengambil left component of the link, contoh [musical\_notation|bar] maka ambil musical\_notation sbg **label**. Kalo cuma [kata] dan katanya ga nunjuk ke page ambigu, ambil ‘kata’ tersebut.

Finally, **labelnya** di MAP MANUAL ke wordnet sense yang cocok. Pake dua orang annotator dan orang ketiga sebagai ‘judge’ annotator. Dihitung kappa dll2 persetujuanya.

and then map these labels to a widely used sense inventory, namely WordNet.

Proses WSD mengintegrasikan **local** and **tropical** feature.

Pertama, text ditokenisasi dan dianotasi dengan POS Tag. Collocation diidentifikasi dengan sliding window approach.

Kedua, ambil fitur local dan topical dari konteks ambiguous wordnya. Yang secara detail terdiri dari:

* Kata yang ambigu (current word), POS tagnya
* Konteks local dari tiga kata kiri dan kanan
* POS tag dari kata2 disekelilingnya
* Verb dan Noun di sebelum dan sesudah kata yang ambigunya

Global konteks diimplementasikan dengan sense-spesific keywords sebagai list dari top 5 words yang muncul minimal 3 kali di konteks defining certain word sense.

we performed a word sense disambiguation experiment on a subset of the ambiguous words used during the SENSEVAL-2 and SENSEVAL-3 evaluations”s

##############################KNOWLEDGE RICH#############################

Resource:

1. Wordnet: konsep wordnet itu direpresentasikan sebagai synonym set (synset) -> set of kata yang punya makna yang mirip.
2. Wikipedia: halaman pada Wikipedia merepresentasikan konsep suatu hal. Misal Soda(Soft Drink), Halaman ini biasanya mengandung hyperlink ke page lain yang relevan seperti Cola, Lemonade, dll.

Proses mapping dari Wikipedia ke Wordnet

Given:

* Set of pages Sensewiki
* Set of pages Sensewn (wordnet)

Sehingga untuk setiap wikipage w merupakan anggota dari Sensewiki dapat:

s jika Sensewn(w) -> set of senses dari lemma dari w di wordnet.

Null jika tidak bisa dikasih link. Untuk established mapping tersebut, cari disambiguation context dari masing-masing Wikipages dan Wordnet. Lalu cari intersection dari context tersebut untuk melakukan mapping.

Nantinya hasil mapping tersebut akan ditambahkan sebagai synset dari wordnet. Misalnya Soft Drink akan terhubung dengan Soda. Dll.

Evaluasi dari hasil mapping dilakukan manual dari list of intersection antara kedua resource. Dites dengan WSD Algorithm dibandingkan dengan manual annotated dataset. Dengan parameter evaluasi:

* Precision (perbandingan dari sense label yang benar dengan label output mapping yang tidak kosong)
* Recall (perbandingan dari sense label yang benar dengan total non empty label dalam gold standard)
* F-Measure biasa (2PR/P+R)

Sebagai baseline, kita menggunakan most frequent Wordnet sense (MFS), dan juga random sense assignment.

Experimen dalam task WSD tersebut menggunakan dua algorithm:

1. Simplified Extended Lesk (ExtLesk):

Based on overlap antara context disamping-samping kata ambigu dengan definisi dari sensenya. Kita juga extend gloss dari sense2 yang ada dengan kata-kata dari gloss synset2nya.

1. Degree Centrality:

Graph based algorithm. Setiap sense s dari target word. Lakukan DFS dari graph wordnet++ dan ambil jalur2 yang menghubungkan s ke sense dari kata lain di konteks tersebut. Pada akhirnya, vertex dengan degree terbesar akan diambil sebagai sense terbaik.

###############Techniques for unsupervised word sense induction################

“The use of word senses in place of surface word forms has been shown to improve performance on many natural language processing tasks such as information ex- traction [Chai and Biermann, 1999], information retrieval [Uzuner et al., 1999], and machine translation [Vickrey et al., 2005]”

Task untuk mengambil sense secara otomatis (WSI) dapat dikonsepkan sebagai permasalahan clustering.

Teknik dari word clustering:

Ubah word menjadi feature vectors. Gunakan algoritma clustering untuk task tersebut (kmeans, buckshot, average link, CBC, dll).

Evaluasi:

Menghitung similarity antara dua synset dari wordnet. Perhitungan similarity tersebut bisa menilai bahwa suatu cluster c, dengan word w di dalamnya, berkorespondensi dengan synset yang di dalamnya terdapat w. Similarity antara w dan synset s tersebut, didefined oleh max similarity antara s dan semua sense dari w. Similarity antara cluster c dengan synset s defined dari rata-rata simarity antara s dengan top-k words dari cluster c. kalau similarity tersebut > threshold, maka c korespondensi dengan s, maka correct word sense.

Extended clustering:

1. Triplet clustering

Dari pada menggunakan word pairs dalam konteks, gunakan triplet.

1. Self term expansion

Buat co-occurrence list of words berdasarkan PMInya, dari list tersebut, words bisa direplace dengan co-related wordnya (diexpand).

1. Clustering of context

“This allows for clustering of local co-occurrences based on context rather than global co-occurrences based on words themselves.” [Michael]

1. Bayesian Approach