Context Specific WSD using Alpha Expansion

BY ARUNAV SANYAL

This report has been compiled as a summary of the work done for Undergraduate Summer Internship in IIT Bombay for the period 21st May 2012 to 18th July 2012. This internship was done in the Centre for Indian Language Technology (CFILT) lab of IIT Bombay.

Report Compiled by:-

Mr Arunav Sanyal
4th year Undergraduate student
CS and IS Department
BITS Pilani K.K. Birla Goa Campus

Project Guide:-

Dr Pushpak Bhattacharya
Professor
CSE Department
IIT Bombay

Acknowledgements:-

I would like to thank my parents for their endless support and encouragement. Special thanks are reserved for Mr Salil Joshi – M. Tech Student of IIT Bombay CSE Department for providing me with the initial assistance in programming the data structures for the problem and with clarification of the basic concepts of Unsupervised WSD. I would like to thank Mr Samiulla Sheikh –M. Tech Student IIT Bombay CSE Department for critically analysing my work and for providing valuable suggestions in correcting flaws in the programs that I developed. I would also like to thank fellow Intern Diptesh Kanoria for some light and fun moments during what was a demanding project.

But most importantly I would like to thank Dr Pushpak Bhattacharya for giving me this opportunity and allowing me to contribute in the field of Natural language processing in my small and limited capacity.

ABSTRACT:

This project attempts an implementation of a new approach to Word Sense disambiguation (assigning meaning to a particular word of a sentence). This problem typically arises in case of polysemous words i.e. words that have more than one meaning and is considered as an A.I Complete problem. My approach to this problem uses an unsupervised method and uses contextual clues in order to facilitate correct assignment of senses. The use of contextual clues to facilitate WSD is polynomially equivalent to solving a Markov Random Field Maximum apriori probability estimation problem. There are many known methods to solve this problem but it is considered as an NP-hard optimization problem. One particularly promising method called Alpha Expansion is used for this project. The Alpha Expansion algorithm is a heuristic algorithm typically used in Computer Vision applications to solve the MRF MAP problem. It reformulates a MAP problem into a graph min-cut problem. This heuristic was modified to be used in WSD and then tested on CFILT Multilingual Corpora. The results are noted and reasons are stated as a possible explanation of observations. Finally possible future methods are conjectured.

Table of Contents:-

1. Introduction
2. Previous work on WSD using Bilingual EM method
3. Markov Random Fields, the MAP problem and the Min-Cut problem
4. The Alpha Expansion algorithm
5. Formulation of WSD Context problem into MRF MAP problem.
6. Empirical observations, results and analysis
7. Conclusions and possible future work
8. Bibliography

Introduction:-

Natural language processing (**NLP**) is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and natural human languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output. A particularly difficult problem or subfield of NLP is Word Sense disambiguation (WSD), which addresses the process of identifying which sense word (i.e. meaning) is used in a sentence, when the word has multiple meanings(i.e. the word is polysemous). The solution to this problem impacts other computer-related applications, such as discourse improving relevance of search engines, anaphora resolution, coherence, inference etc. Many different approaches to WSD have been analysed and executed. These approaches can be broadly classified in the categories of unsupervised and supervised methods. While supervised approaches typically pertain to algorithms that use tagged corpus as a training base, unsupervised algorithms do not have this restriction. Previous work in unsupervised methods included Bilingual EM formulation [1], Dictionary methods [3] etc.

The former method is used as an initial training method on which the Alpha Expansion algorithm is later run on. The Alpha Expansion algorithm would typically use context i.e. the immediate surrounding lexical neighbourhood of a word. The problem can be paraphrased a MRF MAP problem and Alpha Expansion is a heuristic algorithm which tackles it.

Previous work done on WSD using bilingual EM:-

Previous approaches on WSD were generally context insensitive algorithms. One particular example of this is the Bilingual E.M method proposed by Salil et al [1]. This approach makes use of untagged corpora of two languages. One language is called as the source language and the other is called as the target language. These languages help each other gain sense probabilities based on an approach by Mitesh et al [2] which uses a bilingual E.M formulation as an iterative procedure to improve sense probabilities. Sense probabilities mean the probability of one sense of the word occurring given that the word occurs in the untagged corpus. The probabilities are initially assigned uniformly and they are then iteratively improved by the algorithm in accordance with the aforementioned procedure.

The most probable sense of a word was assigned as a sense to all occurrences of the word in the corpus and accuracy was tested with tagged corpora of identical data. The primary flaw in this approach is that no contextual clues were ever taken into account. The reason this is considered as a drawback is that the sense probabilities are highly domain specific and thus context which suggested otherwise was ignored.

The alpha expansion algorithm addresses and solves this problem. However it is important to note that the Bilingual E.M formulation must necessarily precede the Alpha Expansion Algorithm. This is because the base sense probabilities are still needed in order to facilitate creation of the MRF problem. This issue is discussed in further sections.

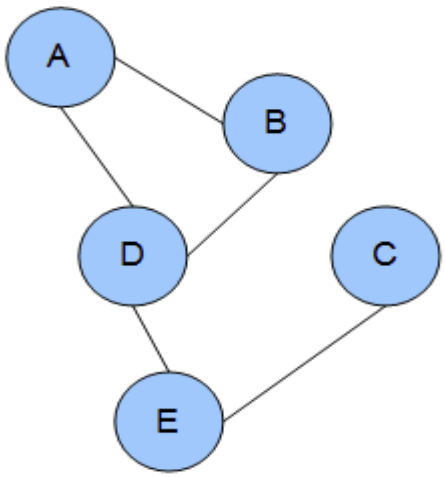
Markov Random fields, the MAP problem and the Min-Cut problem:-

Before we discuss the Alpha Expansion Algorithm, it is crucial to understand the underlying theory behind it.

Markov random field:-

A Markov random field (often abbreviated as MRF), Markov network or undirected graphical model is a set of random variables having a Markov property and is described by an undirected graph. A Markov random field is similar to a Bayesian in its representation of dependencies; the differences being that Bayesian networks are directed and acyclic, whereas Markov networks are undirected and may be cyclic. The MRF contains terms called potentials (which is closely related to the probability of occurrence of a random variable). So higher the potential of a possible occurrence the higher would be the probability of that occurrence. These potentials are for both nodes and edge potentials and they thus termed node and edge potentials respectively.

A Markov property can be typically described as the memory less property of a random variable. This means that the probabilities associated with possible outcomes of a variable depends only on its current state and not on any previous experiments or outcomes.



Consider the image shown above. It is the graphical representation of a MRF

problem. Each node represents a particular random variable and each edge corresponds to potentials of connecting nodes (connecting random variables) occurring together. Note that if node A has n possible outcomes (and corresponding n potentials can be stored in an array of size n) and B has m possible outcomes, then the edge connecting A and B have n*m possible outcomes and the potentials can be stored in a matrix of the same dimension.

Also note that the edge potentials are not related to the node potentials directly.

The MAP problem:-

The maximum apriori estimation problem is the assignment of random variable values that lead to a maximum possible estimation of total summed potentials over the MRF Graph. Both edge and node potentials are taken into consideration for calculating the estimate. This value estimate (random variable assignment values) over all random variables is called as a **label**. A partial labelling is a label that does not cover all variables.

Consider the simple example:-

A--------B is an undirected graph.

A has two labelling possibilities:-
0 -> 0.9 and 1 ->0.4 where the fractions are node potentials

For B:-
0-> 0.8 and 1->0.3

And for edge A-B
0,0 -> 0.7 ; 0,l -> 0.2; 1,0 -> 0.1 ; 1,l -> 0.1
Here the fractions are edge potentials.

Obviously the maximum assignment is 0.9 + 0.8 + 0.7 which is 2.4 and the labelling is A=0 and B=0. This problem was trivially solved because the search space of the problem consisted of only 4 possible labels.

In general the problem is NP-hard (its decision variant is NP-complete). Many

methods exist for solving this problem, some based on linear programming and other optimization techniques [3]. A particular class of methods are based on the min-cut on s-t graphs method. The Alpha Expansion method is the most widely recognized one amongst them.

Min-Cut problem:-

Key to understanding the Alpha Expansion algorithm is to understanding a Min-Cut problem. A minimum cut of a graph is a cut whose cut set has the smallest number of elements (un-weighted case) or smallest sum of weights possible. The min cut set created by the cut is a set of edges whose total summed weight is the minimum of all possible cut sets. There exists many efficient min cut algorithms. This is due to the fact that the min cut problem is equivalent to the well-known Max flow problem which is very easy. We have used the Ford Fulkerson Max Flow estimation algorithm for determining min cuts.

The min cut problem is equivalent to the binary MRF MAP problem. Binary MRFs are discussed in the next section. The min cut thus is a crucial subroutine in the main Alpha Expansion Algorithm.

The Alpha Expansion Algorithm:-

The alpha Expansion algorithm was originally a computer vision heuristic [4]. This algorithm was used in case of non-binary MRFs which was a common feature of many computer vision problems (a binary MRF is one in which each node has a maximum of two possible labels).

The algorithm takes the input of a MRF Graph and gives a complete labelling that corresponds to a MAP solution.

The labelling assignment of a MRF Graph roughly corresponds to the following equation:-

$$E(f) = \sum(D(p)) + \sum(V(p, q))$$

Here E corresponds to the energy (an inverse measure of probability) corresponding to a label and D corresponds to node potentials and V corresponds to edge potentials. p and q are partial labels or assignments over which summations are carried out. Note that Energies are -log(potentials), one should remember that potentials are a direct measure of probability.

There are two requirements for the Alpha Expansion Algorithm:-

1. The edge energies must be symmetric(transpose of edge energy matrix = matrix itself). This condition is optional in general but necessary in case of WSD. This will be discussed later.

2. The edge energies must be sub modular

The condition of sub modularity is actually a generalised version of the condition of being a metric. Mathematically edge energies are sub modular if:-

$$V(a,c) + V(b,b) <= V(a,b) + V(b,c)$$

Over all possible partial labels a, b and c. Here V stands for edge potentials.

The algorithm also requires the energies to be normalized. Normalization means that each edge energy matrix must a minimum value of 0 and each node energy array must also have a minimum energy of 0. The node, edge energies if not normalized must be subtracted with their corresponding array, matrix minimum respectively.

The algorithm first requires creation of a Binary MRF graph from the original MRF graph according to the following equations:-

E'(0) = E(i) //i is the original or initial labelling
E'(1) = E(v) //v is a labelling generated by the main Alpha Expansion algorithm
E'(0,0) = E(i, j) //this is an Edge Energy. i and j are initial labellings
E'(0,1) = E(i, v)
E'(1, 0) = E(v, j)
E'(1,1) = E(v, v)

The energy terms E' are energies corresponding to the binary MRF whereas E terms correspond to the original MRF.

This binary MRF Graph can now be converted into an s-t graph and the min-cut of this graph can be considered as a partial labelling (modified labelling) in the positions corresponding to i and j. The min-cut problem is a restatement of the max-flow problem [4] and has many known solutions.

The following procedure creates an s-t graph from the created binary graph and outputs a min cut edge set. This procedure directly assumes a known min cut finding algorithm. In our case we have used the Ford Fulkerson Max flow algorithm because it gives really good practical run time for small graphs and the min – cut problem can easily be restated as a max – flow problem.

The exact procedure can be viewed conveniently as the following pseudo code:-

```
Procedure MinCut-MAP (
    ϵ    // Singleton and pairwise submodular energy factors
)
1       // Define the energy function
2     for all i
3       ϵ'_i ← ϵ_i
4     Initialize ϵ'_{i,j} to 0 for all i, j
5     for all pairs i < j
6       ϵ'_i(1) ← ϵ'_i(1) + (ϵ_{i,j}(1,0) − ϵ_{i,j}(0,0))
7       ϵ'_j(1) ← ϵ'_j(1) + (ϵ_{i,j}(1,1) − ϵ_{i,j}(1,0))
8       ϵ'_{i,j}(0,1) ← ϵ_{i,j}(1,0) + ϵ_{i,j}(0,1) − ϵ_{i,j}(0,0) − ϵ_{i,j}(1,1)
9
10      // Construct the graph
11    for all i
12      if ϵ'_i(1) > ϵ'_i(0) then
13        E ← E ∪ {(s, z_i)}
14        cost(s, z_i) ← ϵ'_i(1) − ϵ'_i(0)
15      else
16        E ← E ∪ {(z_i, t)}
17        cost(z_i, t) ← ϵ'_i(0) − ϵ'_i(1)
18      for all pairs i < j such that ϵ'_{i,j}(0,1) > 0
19        E ← E ∪ {(z_i, z_j)}
20        cost(z_i, z_j) ← ϵ'_{i,j}(0,1)
21
22      t ← MinCut({z_1, ..., z_n}, E)
23        // MinCut returns t_i = 1 iff z_i ∈ Z_t
24      return t
```

The partial label reassignments are only accepted if leads to a reduction of total energy from the previous assignment. Now this procedure is repeated over all possible labels v and a complete labelling is generated and is defined as one iteration of the algorithm. This new complete label created is said to be one alpha expansion step away from the original label. So each step(iteration) takes the original MRF Graph and an input complete label and creates a new complete label one alpha Expansion step away. This new label is then used as successive input (you can imagine this as a feedback system) and the process repeats itself. The procedure terminates when the following convergence criteria is met – if an input label is equal to the output label after some iteration, the labels converge and the iterations stop. After termination the final output label is a solution to the MAP problem.

The following is the exact pseudo code for the above stated procedure.

```
Procedure Alpha-Expansion (
    ϵ,    // Singleton and pairwise energies
    x     // Some initial assignment
)
1     repeat
2       change ← false
3       for k = 1, ..., K
4         t ← Alpha-Expand(Φ, x, v_k)
5         for i = 1, ..., n
6           if t_i = 1 then
7             x_i ← v_k    // If t_i = 0, x_i doesn't change
8             change ← true
9     until change = false
10    return (x)

Procedure Alpha-Expand (
    ϵ,
    x     // Current assignment
    v     // Expansion label
)
1     Define ϵ' as in equation (13.34)
2     return MinCut-MAP(ϵ')
```

The alpha expansion is a minimization procedure and thus Energies were used instead of potentials.

Note that this procedure is a heuristic with no guarantee of a correct answer. The algorithm runs in practical linear time. However this algorithm has no guarantee of giving the correct answer if the edge energies are non-sub-modular. A similar approach was attempted which is called as the Alpha Beta Swap which swaps two labels at a time. The label swaps are carried out iff(iff is deliberate) the swap results in a better partial label assignment. All possible label combinations are tried and then the final complete labelling is given as an output. This approach was deemed insignificant because output was identical to Alpha Expansion and there was no theoretical bounds on accuracy[4], and also it was practically quadratic in run time. Note that Alpha Beta Swap does not need the edge energies to be Sub modular. This was the primary motive for implementation of this algorithm.

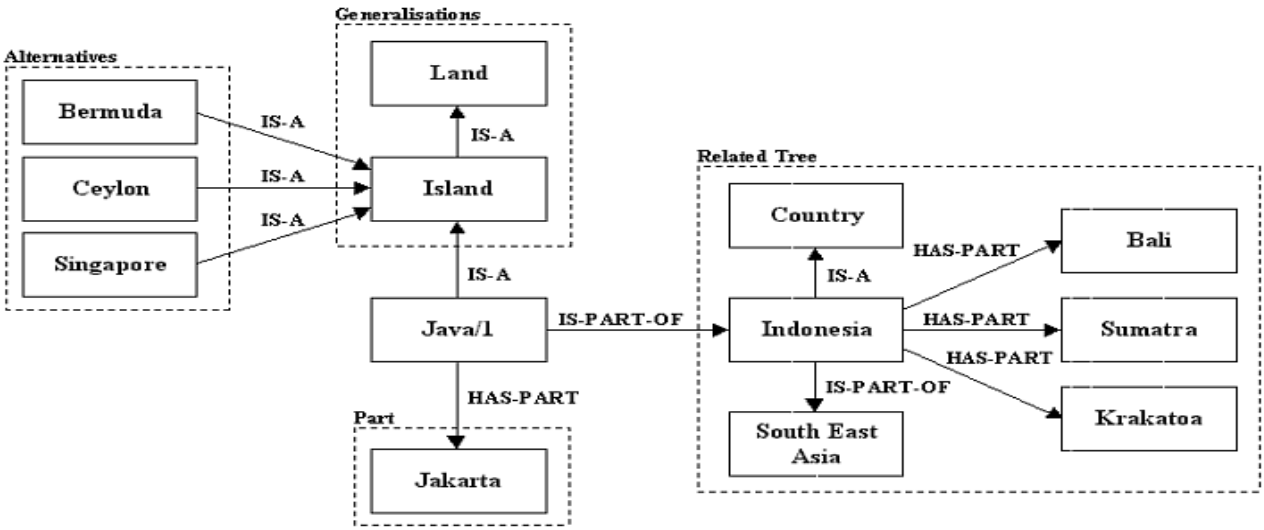
Formulation of WSD Context problem into MRF MAP problem:-

The procedure for converting the WSD Context problem typically involves the creation of an MRF Graph from a given Context. The context in this case was defined as one sentence(delimited by a full stop). All monosemous words(words which have exactly one dictionary sense) were ignored. This typically reduced the problem to an MRF graph consisting only of polysemous words and reduced the computational requirements of the problem. This also ensured creation of small graphs. However it destroyed part of the context. This drawback was discussed later.

Now the question remains of how to assign the edge and node potentials (and hence probabilities).

For node potentials, the approach of Bilingual E.M was used, which was discussed in section 2. The word sense probabilities generated by this algorithm were directly taken as potential values for a node. Each word was deemed as a node and its possible senses possible partial labels for the node. The node potentials were normalized and converted to energies (-log(potentials)).

For edge potentials a radically different technique was used. The technique used was called as the conceptual distance – which is a measure of semantic similarity of words based on the Wordnet. The Wordnet [5] is a lexical database which categorizes words according to synonym sets (synsets) and relationships between them are established (hypernym- i.e. is-a type relationship, meronym i.e. - part-of type relationship etc.). This forms a hierarchy of words which looks like tree like structure with the most obscure or generic words in meaning in top of the tree to most particular words in meaning in bottom of the tree. The wordnet is categorized into verbs, nouns, adjectives and adverbs and the semantic relationships differ for these main types.



The above diagram is a part of a noun part of the wordnet which is horizontally arranged.

Mathematically the conceptual distance is[6]:-

$$DA(n1, n2) = d(n1, n2) / f(L)$$

Here n1 and n2 are two nouns and d is the hypernymy distance between them(i.e. the number of distinct hypernym relationships separating them). In graphical terms it is the graph traversal distance between the two nouns in the hypernymy hierarchy of the wordnet(i.e. the minimum node to node distance possible in a graph). The function f is a density measure. Roughly speaking f is higher if the nouns lie in a region L and the region is densely populated(more number of semantic relationship links in that region). The region L is defined as a horizontal subsection of the hypernymy tree. For example, the top three levels of nodes can be considered as a region. Similar measures can be defined for other POS categories with hypernymy replaced by other relationship categories (e.g. troponymy – i.e. manner- of type relationship in case of verbs).

This measure creates a value which is used as a potential value which is later normalized and converted to Energy values.

The earlier question about symmetry of MRF Edge potentials can now be answered. Technically the transpose of the matrix corresponding to the edge potentials can be

viewed as a rearrangement of labels. Thus in the context of conceptual distance, it is a rearrangement of words given as an input to the DA() function. The conceptual distance is invariant to the word order as only the numerator arguments are altered which in turn is a distance function. In other words the distance between word1 and word2 in a Wordnet hierarchy is same as the distance between word2 and word1.

Empirical observations, results and analysis:-

All work was carried out in Java and the cfilt server was used as a work station. The following output is a demonstration of the correctness of the Alpha Expansion Algorithm was sub modular graphs(graph in which all edges are sub modular):-

Consider the graph as follows(This is direct output from code):-

Input
3 2
0 2 0.1 0.8
1 2 0.4 0.1
2 2 0.6 0.3
0 1 0.1 0.2 0.9 0.3
0 2 0.1 0.7 0.8 0.6

Output
[1 0 0] - 3.5

Explanation :- The first two numbers of the input are no of nodes and no of edges. The next 3 lines give node number, max label size and node potentials in order. The last two lines are edge potentials along with node numbers of the nodes forming the edge. The output corresponds to the MAP assignment, maximizing potential sum with the sum being 3.5. The algorithm took one iteration in which the labelling changed and converged at second step, thus it took a total of two iterations.

The correctness was verified by a brute force algorithm as well. Many more manual examples on sub modular graphs were tried and all gave correct output. As of writing this report, the algorithm never failed on sub modular graphs.

The following is the final output statistics for the Marathi Corpora using the Hindi Corpora as the source language for the initial EM training. Tests were done on Health and Tourism domain:-

	POS wise and overall in percentages				
Domain	Noun	Adjective	Adverb	Verb	Overall
Health	46.08	39.26	53.04	58.42	50.72
Tourism	48.12	19.14	61.11	65.85	55.86

Convergence is always observed in practice for all corpora. Most incorrectly tagged words that were manually checked in HEALTH domain turned out to be nouns. This was because nouns constitute the sheer bulk of the words in HEALTH domain and they have a low correct-tagging percentage as well.

The output is less accurate than the Bilingual E.M Formulation[1](approx. 56 % in both domains). The following could be the reasons for this:-

1. Exceedingly high number of non-sub modular edges were encountered in the corpora. The alpha expansion algorithm gives bad results in such cases. This is conjectured to be the principal reason behind the inefficiency.

2. Ignoring monosemous words. Ignoring these words could mean critical context clues were lost. This could be a major reason for the inaccuracies of this method.

3. Use of conceptual distance could lead to incorrect output. The wordnet based method and the bilingual method have nothing in common. It could be that the edge energies are incorrectly assumed leading to erroneous results.

Although results weren't up to mark since benchmark was against a top notch algorithm [1], it nevertheless beats many state of the art algorithms for unsupervised WSD [1].

I also implemented a partial brute force method (brute force on smaller graphs) in order to check for correctness and to ascertain the theoretical maximum efficiency possible for this formulation of MRF MAP problem. The algorithm gave better results but not comparable to supervised techniques. It too gave a pessimistic bound of approx. 59% for both domains.

Conclusions and possible future work:-

Although the algorithm gave pessimistic results, it was a great learning experience. The algorithm implementation was perfect. On a personal note, I gained immense experience in JAVA and acquired the ability to work in large projects. More specifically I learned how to understand data structures created by other students and how to use them effectively. Also, I learned to discipline myself mostly because I was unsupervised :) . The project gave me good crash course on WSD and A.I research in general. In the end it gave me a great opportunity to utilize my vacation constructively(which would otherwise have been spent watching South Park)

Other approaches like the QBPO(Quadratic Boolean Optimization)[7] algorithm could be used in order to avoid the non-sub modularity problem. More robust semantic measures could also be used. Although a complete Brute force Algorithm is out of the question, a modified Alpha Expansion algorithm could be tried which uses brute force in non-sub modular cases. Another approach that was considered during the discussions was direct application of contextual clues during the E.M formulation [1] but this method is still conjectural but could become significant future work.

In the end it was great but a lot more could be done.

Bibliography:-

1. It Takes Two to Tango: A Bilingual Unsupervised Approach for Estimating Sense Distributions Using Expectation Maximization. Mitesh M. Khapra Salil Joshi Pushpak Bhattacharyya Department Of Computer Science and Engineering, IIT Bombay, Powai, Mumbai, 400076.

2. Mitesh M. Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2009. Projecting parameters for multilingual word sense disambiguation. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 459–467, Singapore, August. Association for Computational Linguistics

3. Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet by Satanjeev Banerjee December 2002 Submitted in partial fulfilment of the requirements for the degree of Master of Science under the instruction of Dr Ted Pedersen

4. Probabilistic Graphical Models: Principles and Techniques By Daphne Koller, Nir Friedman

5. Introduction to WordNet: An On-line Lexical Database George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller (Revised August 1993)

6. Measuring Conceptual Distance Using WordNet: The Design of a Metric for Measuring Semantic Similarity William D. Lewis University of Arizona

7. Pseudo-Boolean Optimization Endre Boros† and Peter L. Hammer† October 15, 2001