Volume 5, No. 1, January 2014
Journal of Global Research in Computer Science
ISSN-2229-371X
REVIEW ARTICLE
Available Online at www.jgrcs.info

CORPUS ALIGNMENT FOR WORD SENSE DISAMBIGUATION

Shweta Vikram
Computer Science, Banasthali Vidyapith, Jaipur, Rajasthan, India
shwetavikram.2009@rediffmail.com

Abstract: Machine translation convert one language to another language. Anusaaraka is a machine translation, which is an English to Indian language accessing software. Anusaaraka is a Natural Language Processing (NLP) Research and Development project undertaken by Chinmaya International Foundation (CIF). When any machine do that work they need big parallel corpus that can help for making some rules and disambiguate many senses. It is following hybrid approach but we are working on rule based approach. For this approach we needed big parallel aligned corpus. In this paper we discuss how we collect parallel corpus with the help of some shell scripts, some programs, some tool kit and other things.
Keyword: Corpus, alignment, Champollion tool kit, English files and Hindi files.

INTRODUCTION

Parallel corpus has English sentences and corresponding Hindi sentences. It is not necessary corpus is always parallel corpus it may or may not be. In this paper we are talking about English and Hindi corpus. If any corpus available in only English or Hindi then it is not necessary this is parallel corpus. Gyanidhi is a parallel English-Hindi corpus. Corpus is a singular and corpora are a plural for word corpus.

Whenever machine translation translates English to Hindi language then this machine required big English-Hindi parallel corpus. Because the help of this corpus we analyzed the relation between word. Today the collection of parallel corpus is very big problem because manual translation takes much more time.

CORPUS ALIGNMENT

Corpus collection

In collection of corpus the main work is collect the parallel corpus. Collect both files English as well as Hindi translation files. If we have any one file (English or Hindi) then we search for other translated file (Hindi or English). Many website having both languages. Some are:

www.loksabha.nic. in
www.loksabhahindi.nic.in
www.rajyasabhahindi.com
www.hindisamay.com
www.gyanidhi.com etc.
www.rajyasabhahindi.com

Loksabha, loksabhahindi, rajyasabha, rajyasabhahindi websites have many more English files and that Hindi files. In hindisamay.com have only Hindi files but this corpus is very big Hindi corpus. Gyanidhi have both files that means it have parallel corpus.

If we have English file then we searching for its corresponding Hindi translation file, if we get that file then those files are useful for us otherwise not. Whenever we are not getting parallel files then our time, effort and money are waste.

Corpus alignment

Alignment is play important role in preparing parallel corpus. In Corpus alignment firstly we are collecting English file and corresponding Hindi file or vice versa. Our parallel corpus should be in txt format, because our machine translation software Anusaaraka read only txt files.

Some steps for corpus alignment

1. We collect English file and corresponding Hindi translation file. If both files are in txt format it is good for us otherwise we convert into txt format with the help of shell script. We are converting:

html to txt
pdf to txt

English files are easily convert one form to another but Hindi files are not easily converted because they need some mapping scripts for that files (Hindi).

2. After getting txt format we performed ' one line per sentence' operation on both files that means every line started at new line.

Use of Champollion toolkit

History

The first attempt to automatically align parallel text was (Gale and Church 1991), which is based on the idea that long sentences will be translated into long sentences and short sentences into short ones. A probabilistic score is assigned to each proposed correspondence of sentences, based on the scaled difference of lengths of the two sentences and the variance of this difference. The probabilistic score is used in

a dynamic programming framework to find the maximum likelihood alignment of sentences. The length based In addition to the length based approach and length and lexicon hybrid approach, there are a few other approaches in the literature. (Chen 1996) builds a sentence-based translation model and fined the alignment with the highest probability given the model. (Melamed 1999) first finds token correspondences between the source text and its translation by using a pattern recognition method. These token correspondences are used in conjunction with segment boundary information to find sentence correspondences

Champollion Tool Kit (CTK) is a tool kit aiming to provide ready-to-use parallel text sentence alignment tools for as many language pairs as possible. Built around the LDC Champollion sentence aligner kernel, the tool kit provides essential components required for accurate sentence alignment, including sentence breakers, stemmers, pre-processing scripts, dictionaries (if possible), post-processing scripts etc. Currently, CTK includes tools to align English text with Arabic, Chinese, and Hindi translations. It can be easily expanded to other language pairs.CTK welcomes contributions from other researchers. CTK is written in perl.

Working of Champollion Toolkit

The input files for both sides should be one segment (sentence) per line. The output (alignment file) looks like the following:

omitted <=> 1
omitted <=> 2
omitted <=> 3
1 <=> 4
2 <=> 5
3 <=> 6

Where each language1/language2 sentence ids may contain up to four sentence ids delimited by commas, it also can be "omitted" indicating no translation was found. The sentence ids start at 1.Languages CTK v1.2 supports three language pairs:

1. Toolkit is parallel text is one of the most valuable resources for development of statistical machine translation systems and other NLP applications.

2. It takes two input files english.txt and hindi.txt, and generates one align_output.txt file.

3. These file shows the English sentences are corresponding Hindi sentences are or not.

Champollion differs from other sentence aligners in two ways. First, it assumes a noisy input, i.e. that a large percentage of alignments will not be one to one alignments, and that the number of deletions and insertions will be significant. The assumption is against declaring a match in the absence of lexical evidence. Non-lexical measures, such as sentence length information –which are often unreliable when dealing with noisy data – can and should still be used, but they should only play a supporting role when lexical evidence is present. Second, Champollion differs from other lexicon-based approaches in assigning weights to translated words. Translation lexicons usually help sentence aligners in the following way: first, translated words are identified by using entries from a translation lexicon; second, statistics of translated words are then used to identify sentence correspondences. In most existing sentence alignment approach works remarkably well on language pairs with high length correlation.

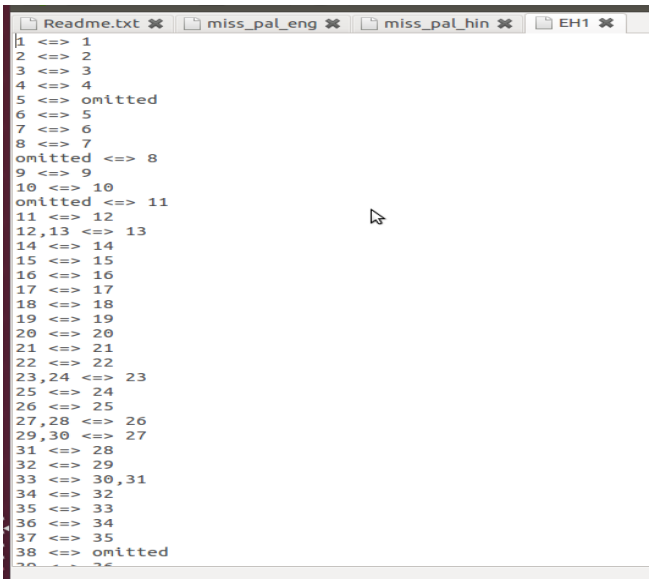

Figure1. Champollion tool kit Output.

Translated words are treated equally, i.e. translated word pairs are assigned equal weight when deciding sentence correspondences. Should these translated word pairs have an equal say about whether two sentences are translations of each other? Probably not. For example, assume that we have the following sentence pair in a one file of loksabha:

English: Starred Questions Starred Question No. 41 was orally answered.
Hindi in 'wx' notation: wArAMkiwa praSna wArAMkiwa praSna saMKyA 41 kA mOKika uwwara xiyA gayA.

English: In pursuance of the aforesaid resolutions, the matter was considered by the Rules Committee (Eleventh Lok Sabha).
Hindi in 'wx' notation: uparyukwa saMkalpa ke anusaraNa meM niyama samiwi (gyArahavIM loka saBA) xvArA isa mAmale para vicAra kiyA gayA WA.

English: The Rules Committee in their Second Report laid on the Table of the House on 6th March, 1997 recommended that a Committeefor the purpose may be constituted.
Hindi in 'wx' notation: niyama samiwi ne 6 mArca, 1997 ko saBA patala para raKe gae apane xUsare prawivexana meM siPAriSa kI ki isa prayojana ke lie eka samiwi gaTiwa kI jAe.

English: Accordingly the Committee on Empowerment of Women was constituted on 29th April, 1997.
Hindi in 'wx' notation: waxanusAra, 29 aprEla, 1997 ko mahilAoM ko SakwiyAM praxAna karane saMbaMXI samiwi gaTiwa kI gaI.

English: Appointment of Chairman The Chairman of the Committee is appointed by the Speaker from amongst its

Members.
Hindi in 'wx' notation: saBApawi kI niyukwi samiwi kA saBApawi aXyakRa xvArA samiwi ke saxasyoM meM se niyukwa kiyA jAwA hE.

English: The Committee may also examine any other subjects/matters for improving the status/condition of women which come within its purview.
Hindi in 'wx' notation: samiwi mahilAoM kI sWiwi/xaSA meM suXAra ke lie kisI anya viRaya/mAmale para BI vicAra kara sakawI hE jo isake kRewrAXikAra meM Awe hoM.

Here we discuss first sentence the translation pair (41, 41) is much stronger evidence than (Starred, wArAMkiwa) that the two sentences are a match, simply because "Starred" and "wArAMkiwa" appear much more function to compute the similarity between any two segments, each of which consists of one or more sentences. There is a penalty associated with alignments other than 1-1 alignment. The penalty is determined empirically. Sentences with a mismatching length are also penalized. Champollion then uses a dynamic programming method to find the optimal alignment which maximizes the similarity of the source text and the translation.

When we are getting this align_output.txt file then we are use this file in our another program 'align-eng-hnd'. This program generates the output which contains parallel sentences of both the files. This program takes three input files align_output.txt, Hindi.txt and English.txt then generate one output file (output.txt).

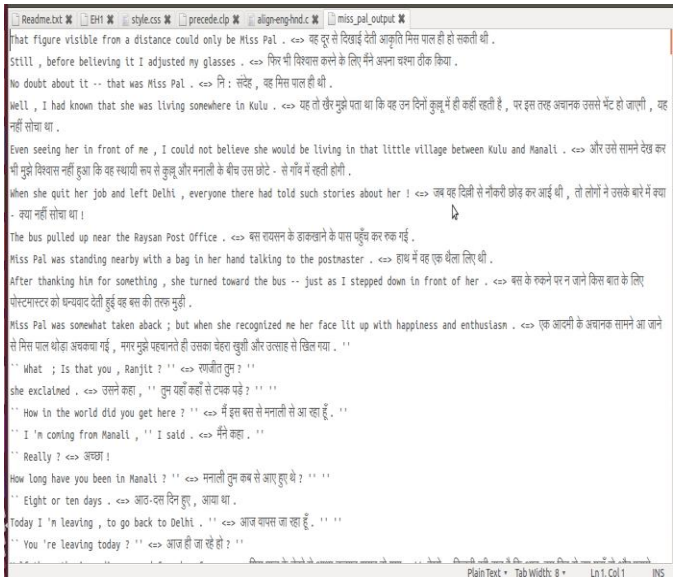

Figure2. Show the output.txt file

Splitting English Hindi files

When we getting output.txt file it contain both translation English as well as Hindi translation sowed split into two files one file contain only English sentences and another file contain only Hindi sentences. Finally we got parallel aligned corpus.

IMPROVE THE CHAMPOLLION DICTIONARY

When we are want to get better result than we add more words and phrasal word in dictionary. So we run parallel corpus on machine which generate phrasal verb. These phrasal verbs add on Champollion dictionary and next time when we run parallel corpus then it give improved output.

Some phrasal verb we get and we add on Champollion dictionary. Here we show the some phrasal verb.

twenty-first_day <> ikkisawe din
two_Total_Result <> do_sakal_parinam

two_Total_Result_Display <> do_Sakai_parinam_suchak

two_consecutive <> do_niranntar

two_consecutive_days <> do_nirannatar_din
urgent_and <> awilambiy_aur

urgent_and_important <> awilambniy_air_mahatawpuna

urgent_public_importance <> awilambniy_look_mahatwa
various_Corporations <> wibhinn_nigamo
wherever_necessary <> jaha_aawashyak

MACHINE TRANSLATION

Anusaaraka

Anusaaraka is English to Indian language accessing (translation) software, which employs algorithms derived from Panini's Ashtadhyayi (Grammar rules). The software is being developed by the Chinmaya International Foundation at the International Institute of Information Technology, Hyderabad (IIIT-H) and the University of Hyderabad (Department of Sanskrit Studies). Anusaaraka is viewed as the fusion of traditional Indian shastras and advanced modern technologies.

Approaches Used

For solving our objective we are using some approaches in our research based project. Various types of approaches of "Word Sense Disambiguation "are:-

a) Rule based approach
b) Machine learning approach
c) Dictionary based approach
d) Hybrid approach

a) ***Rule based approach***
Making some rules for a particular English word which can cover all the Hindi senses with respect to the context.

b) ***Machine learning approach***
This approach want big corpus for working. These approaches find out the nature of sentences and also analyze those sentences and then ready for translation. Machine learning approach divided into three parts:

1. ***Supervised approach***
In supervised learning, the model defines the effect one set of observations, called inputs, has on another set of observations, called outputs. In other words, the inputs are assumed to be at the beginning and outputs at

the end of the causal chain. The models can include mediating variables between the inputs and outputs.

2. ***Unsupervised learning***

In unsupervised learning, all the observations are assumed to be caused by latent variables, that is, the observations are assumed to be at the end of the causal chain. In practice, models for supervised learning often leave the probability for inputs undefined. This model is not needed as long as the inputs are available, but if some of the input values are missing, it is not possible to infer anything about the outputs. If the inputs are also modeled, then missing inputs cause no problem since they can be considered latent variables as in unsupervised learning.

3. ***Semi supervised approach***

The bootstrapping approach starts from a small amount of seed data for each word: either manually-tagged training examples or a small number of surefire decision rules .

Other semi-supervised techniques use large quantities of untagged corpora to provide co-occurrence information that supplements the tagged corpora. These techniques have the potential to help in the adaptation of supervised models to different domains.

c***) Dictionary based approach***

In this many dictionaries are used for translations. Such as for physics related subject required physics dictionary and similarly for other subjects.

d) Hybrid approach

It is the combination of all approaches and follows all approaches when its wants.

Now we have aligned parallel text corpus so run this on anusaaraka machine translation software and this generate layered output and also we get machine translation. Now comparing the Hindi translation and anusaaraka Hindi

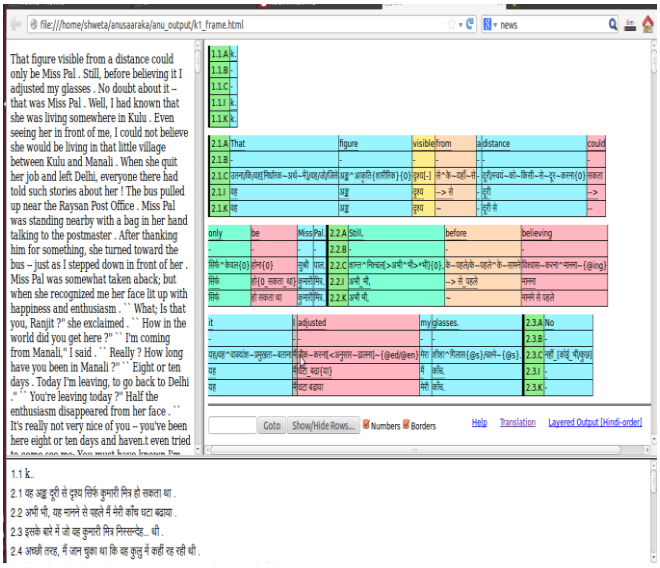

Figure 3.Anusaaraka output

Translation. If both are same and give same meaning of each word then it is good for our software and that means this time no need to new rule for those words. But if any words not have the same meaning in particular context then we will make the rule for that word.

ADVANTAGE, DISADVANTAGE AND APPLICATIONS

Problem in manual translation

When we do anything is manually then it takes time. When we think about manual translation then we need more people, more time and more money. But one thing is very good in manual translation is it give always correct translation or aligned corpus.

Advantages

1. It generates more aligned sentences within few minutes. So it saves the time.
2. Some people do this alignment so money will be saved.
3. Parallel corpus alignment are work done by computer so many extra effort (such as thinking about translation) will be save.

Drawbacks

1. Champollion does not give 100 percent guarantee for those sentences are correct aligned.
2. Some time it gave few outputs but it take more no of sentences. For example if we gave 135 sentences in input file then it gave only 91 aligned sentences, because some sentences are omitted.

Applications

1. We can use aligned corpus in many translation software.
2. This aligned corpus is very useful for case base reasoning (CBR).

Result and discussion

We align many english-hindi files here we show the result.

Table 1. Result of Parallel corpus alignment.

Corpus	No. of sentences in English files	No. of sentences in Hindi files
Short story	1624	1624
Lok shabha	8082	8082
Other	3245	3245

This result show the no of aligned sentences in hindi English and Hindi files. Here we take nine short story written by hindi writer Munshi Premchand and its english translation by Arvind Gupta. We also take many loksabha files and some others files.

CONCLUSION

We are discussing the entire think with respect to the parallel corpus alignment. If we follow manual translation it takes much more time so we not reach our goal in limited time. But we follow the other method which is discussed then we collect many more parallel aligned corpus and reach our goal as soon as possible. With the help of this our aim will be

fulfill and very soon everything will available in Hindi language.

ACKNOWLEDGEMENTS

All praise, loud, and honor to Budhha for His amazing grace and guidance. It is prof. Vineet Chaitanya of IIIT Hyderabad who unraveled my talent and dormant potential of writing a paper of this kind. Spacial thanks to him. I am indebted to Dr. Dipti Mishra Sharma and Sukhada mam for their contribution and suggestion. I am very Thankful of all members of Anusaaraka team, my teachers, my family and my friends.

REFERENCES

[1] Speech and Language Processing an Introduction to Natural
Language Processing, Computational Lingustics, and Speech
Recognition, by Denial Jurasky and James H. Martin.
[2] www.anusaaraka.iiit.ac.in

[3] www.wikipedia.org

[4] www.Sanskrit.inria.fr/DATA/wx.html

[5] www.loksabha.com
[6] www.loksabhahindi.com
[7] www.rajyasabhahindi.com
[8] www.hindisamay.com
[9] www.gyanidhi.com etc.
[10] www.rajyasabhahindi.com

