The SENSEVAL–3 Multilingual English–Hindi Lexical Sample Task

Timothy Chklovski
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
timc@isi.edu

Rada Mihalcea
Department of Computer Science
University of North Texas
Dallas, TX 76203
rada@cs.unt.edu

Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN 55812
tpederse@d.umn.edu

Amruta Purandare
Department of Computer Science
University of Minnesota
Duluth, MN 55812
pura0010@d.umn.edu

Abstract

This paper describes the English–Hindi Multilingual lexical sample task in SENSEVAL–3. Rather than tagging an English word with a sense from an English dictionary, this task seeks to assign the most appropriate Hindi translation to an ambiguous target word. Training data was solicited via the Open Mind Word Expert (OMWE) from Web users who are fluent in English and Hindi.

1 Introduction

The goal of the MultiLingual lexical sample task is to create a framework for the evaluation of systems that perform Machine Translation, with a focus on the translation of ambiguous words. The task is very similar to the lexical sample task, except that rather than using the sense inventory from a dictionary we follow the suggestion of (Resnik and Yarowsky, 1999) and use the translations of the target words into a second language. In this task for SENSEVAL-3, the contexts are in English, and the "sense tags" for the English target words are their translations in Hindi.

This paper outlines some of the major issues that arose in the creation of this task, and then describes the participating systems and summarizes their results.

2 Open Mind Word Expert

The annotated corpus required for this task was built using the Open Mind Word Expert system (Chklovski and Mihalcea, 2002), adapted for multilingual annotations [1].

To overcome the current lack of tagged data and the limitations imposed by the creation of such data using trained lexicographers, the Open Mind Word Expert system enables the collection of semantically annotated corpora over the Web. Tagged examples are collected using a Web-based application that allows contributors to annotate words with their meanings.

The tagging exercise proceeds as follows. For each target word the system extracts a set of sentences from a large textual corpus. These examples are presented to the contributors, together with all possible translations for the given target word. Users are asked to select the most appropriate translation for the target word in each sentence. The selection is made using check-boxes, which list all possible translations, plus two additional choices, "unclear" and "none of the above." Although users are encouraged to select only one translation per word, the selection of two or more translations is also possible. The results of the classification submitted by other users are not presented to avoid artificial biases.

3 Sense Inventory Representation

The sense inventory used in this task is the set of Hindi translations associated with the English words in our lexical sample. Selecting an appropriate English-Hindi dictionary was a major decision early in the task, and it raised a number of interesting issues.

We were unable to locate any machine readable or electronic versions of English-Hindi dictionaries, so it became apparent that we would need to manually enter the Hindi translations from printed materials. We briefly considered the use of Optical Character Recognition (OCR), but found that our available tools did not support Hindi. Even after deciding to enter the Hindi translations manually, it wasn't clear how those words should be encoded. Hindi is usually represented in Devanagari script, which has a large number of possible encodings and no clear standard has emerged as yet.

[1] Multilingual Open Mind Word Expert can be accessed at http://teach-computers.org/word-expert/english-hindi

SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

We decided that Romanized or transliterated Hindi text would be the the most portable encoding, since it can be represented in standard ASCII text. However, it turned out that the number of English–Hindi bilingual dictionaries is much less than the number of Hindi–English, and the number that use transliterated text is smaller still.

Still, we located one promising candidate, the English–Hindi Hippocrene Dictionary (Raker and Shukla, 1996), which represents Hindi in a transliterated form. However, we found that many English words only had two or three translations, making it too coarse grained for our purposes[2].

In the end we selected the Chambers English–Hindi dictionary (Awasthi, 1997), which is a high quality bilingual dictionary that uses Devanagari script. We identified 41 English words from the Chambers dictionary to make up our lexical sample. Then one of the task organizers, who is fluent in English and Hindi, manually transliterated the approximately 500 Hindi translations of the 41 English words in our lexical sample from the Chambers dictionary into the ITRANS format (http://www.aczone.com/itrans/). ITRANS software was used to generate Unicode for display in the OMWE interfaces, although the sense tags used in the task data are the Hindi translations in transliterated form.

4 Training and Test Data

The MultiLingual lexical sample is made up of 41 words: 18 nouns, 15 verbs, and 8 adjectives. This sample includes English words that have varying degrees of polysemy as reflected in the number of possible Hindi translations, which range from a low of 3 to a high of 39.

Text samples made up of several hundred instances for each of 31 of the 41 words were drawn from the British National Corpus, while samples for the other 10 words came from the SENSEVAL-2 English lexical sample data. The BNC data is in a "raw" text form, where the part of speech tags have been removed. However, the SENSEVAL-2 data includes the English sense–tags as determined by human taggers.

After gathering the instances for each word in the lexical sample, we tokenized each instance and removed those that contain collocations of the target word. For example, the training/test instances for arm.n do not include examples for *contact arm*, *pickup arm*, etc., but only examples that refer to *arm* as a single lexical unit (not part of a collocation). In our experience, disambiguation accuracy on collocations of this sort is close to perfect, and we aimed to concentrate the annotation effort on the more difficult cases.

The data was then annotated with Hindi translations by web volunteers using the Open Mind Word Expert (bilingual edition). At various points in time we offered gift certificates as a prize for the most productive tagger in a given day, in order to spur participation. A total of 40 volunteers contributed to this task.

To create the test data we collected two independent tags per instance, and then discarded any instances where the taggers disagreed. Thus, each instance that remains in the test data has complete agreement between two taggers. For the training data, we only collected one tag per instance, and therefore this data may be noisy. Participating systems could choose to apply their own filtering methods to identify and remove the less reliably annotated examples.

After tagging by the Web volunteers, there were two data sets provided to task participants: one where the English sense of the target word is unknown, and another where it is known in both the training and test data. These are referred to as the translation only (t) data and the translation and sense (ts) data, respectively. The t data is made up of instances drawn from the BNC as described above, while the ts data is made up of the instances from SENSEVAL-2. Evaluations were run separately for each of these two data sets, which we refer to as the t and ts subtasks.

The t data contains 31 ambiguous words: 15 nouns, 10 verbs, and 6 adjectives. The ts data contains 10 ambiguous words: 3 nouns, 5 verbs, and 2 adjectives, all of which have been used in the English lexical sample task of SENSEVAL-2. These words, the number of possible translations, and the number of training and test instances are shown in Table 1. The total number of training instances in the two sub-tasks is 10,449, and the total number of test instances is 1,535.

5 Participating Systems

Five teams participated in the t subtask, submitting a total of eight systems. Three teams (a subset of those five) participated in the ts subtask, submitting a total of five systems. All submitted systems employed supervised learning, using the training examples provided. Some teams used additional resources as noted in the more detailed descriptions

[2]We have made available transcriptions of the entries for approximately 70 Hippocrene nouns, verbs, and adjectives at http://www.d.umn.edu/~pura0010/hindi.html, although these were not used in this task.

Table 1: Target words in the SENSEVAL-3 English-Hindi task

Lexical Unit	Translations	Train	Test	Lexical Unit	Translations	Train	Test	Lexical Unit	Translations	Train	Test
TRANSLATION ONLY (T–DATA)											
band.n	8	224	91	bank.n	21	332	52	case.n	13	348	42
different.a	4	320	25	eat.v	3	271	48	field.n	14	300	100
glass.n	8	379	13	hot.a	18	348	32	line.n	39	360	11
note.v	11	220	12	operate.v	9	280	50	paper.n	8	264	73
plan.n	8	210	35	produce.v	7	265	67	rest.v	14	172	10
rule.v	8	160	18	shape.n	8	320	32	sharp.a	16	248	48
smell.v	5	210	17	solid.a	16	327	37	substantial.a	15	250	100
suspend.v	4	370	28	table.n	21	378	16	talk.v	6	341	35
taste.n	6	350	40	terrible.a	4	200	99	tour.n	5	240	9
vision.n	14	318	20	volume.n	9	309	54	watch.v	10	300	100
way.n	16	331	22					TOTAL	348	8945	1336
TRANSLATION AND SENSE ONLY (TS–DATA)											
bar.n	19	278	39	begin.v	6	360	15	channel.n	6	92	16
green.a	9	175	26	nature.n	15	71	14	play.v	14	152	10
simple.a	9	166	19	treat.v	7	100	32	wash.v	16	10	11
work.v	24	100	17					TOTAL	125	1504	199

below.

5.1 NUS

The NUS team from the National University of Singapore participated in both the t and ts subtasks. The t system (nusmlst) uses a combination of knowledge sources as features, and the Support Vector Machine (SVM) learning algorithm. The knowledge sources used include part of speech of neighboring words, single words in the surrounding context, local collocations, and syntactic relations. The ts system (nusmlsts) does the same, but adds the English sense of the target word as a knowledge source.

5.2 LIA-LIDILEM

The LIA-LIDILEM team from the Université d' Avignon and the Université Stendahl Grenoble had two systems which participated in both the t and ts subtasks. In the ts subtask, only the English sense tags were used, not the Hindi translations.

The FL-MIX system uses a combination of three probabilistic models, which compute the most probable sense given a six word window of context. The three models are a Poisson model, a Semantic Classification Tree model, and a K nearest neighbors search model. This system also used a part of speech tagger and a lemmatizer.

The FC-MIX system is the same as the FL-MIX system, but replaces context words by more general synonym–like classes computed from a word aligned English–French corpus which number approximately 850,000 words in each language.

5.3 HKUST

The HKUST team from the Hong Kong University of Science and Technology had three systems that participated in both the t and ts subtasks

The HKUST_me_t and HKUST_me_ts systems are maximum entropy classifiers. The HKUST_comb_t and HKUST_comb_ts systems are voted classifiers that combine a new Kernel PCA model with a maximum entropy model and a boosting–based model. The HKUST_comb2_t and HKUST_comb2_ts are voted classifiers that combine a new Kernel PCA model with a maximum entropy model, a boosting–based model, and a Naive Bayesian model.

5.4 UMD

The UMD team from the University of Maryland entered (UMD–SST) in the t task. UMD–SST is a supervised sense tagger based on the Support Vector Machine learning algorithm, and is described more fully in (Cabezas et al., 2001).

5.5 Duluth

The Duluth team from the University of Minnesota, Duluth had one system (Duluth-ELSS) that participated in the t task. This system is an ensemble of three bagged decision trees, each based on a different type of lexical feature. This system was known as Duluth3 in SENSEVAL-2, and it is described more fully in (Pedersen, 2001).

6 Results

All systems attempted all of the test instances, so precision and recall are identical, hence we report

Table 2: t Subtask Results

System	Accuracy
nusmlst	63.4
HKUST_comb_t	62.0
HKUST_comb2_t	61.4
HKUST_me_t	60.6
FL-MIX	60.3
FC-MIX	60.3
UMD-SST	59.4
Duluth-ELSS	58.2
Baseline (majority)	51.9

Table 3: ts Subtask Results

System	Accuracy
nusmlsts	67.3
FL-MIX	64.1
FC-MIX	64.1
HKUST_comb_ts	63.8
HKUST_comb2_ts	63.8
HKUST_me_ts	60.8
Baseline (majority)	55.8

the single Accuracy figure. Tables 2 and 3 show results for the t and ts subtasks, respectively.

We note that the participating systems all exceeded the baseline (majority) classifier by some margin, suggesting that the sense distinctions made by the translations are clear and provide sufficient information for supervised methods to learn effective classifiers.

Interestingly, the average results on the ts data are higher than the average results on the t data, which suggests that sense information is likely to be helpful for the task of targeted word translation. Additional investigations are however required to draw some final conclusions.

7 Conclusion

The Multilingual Lexical Sample task in SENSEVAL-3 featured English ambiguous words that were to be tagged with their most appropriate Hindi translation. The objective of this task is to determine feasibility of translating words of various degrees of polysemy, focusing on translation of specific lexical items. The results of five teams that participated in this event tentatively suggest that machine learning techniques can significantly improve over the most frequent sense baseline. Additionally, this task has highlighted creation of testing and training data by leveraging the knowledge of bilingual Web volunteers. The training and test data sets used in this exercise are available online from http://www.senseval.org and http://teach-computers.org.

Acknowledgments

Many thanks to all those who contributed to the Multilingual Open Mind Word Expert project, making this task possible. We are also grateful to all the participants in this task, for their hard work and involvement in this evaluation exercise. Without them, all these comparative analyses would not be possible.

We are particularly grateful to a research grant from the University of North Texas that provided the funding for contributor prizes, and to the National Science Foundation for their support of Amruta Purandare under a Faculty Early CAREER Development Award (#0092784).

References

S. Awasthi, editor. 1997. *Chambers English–Hindi Dictionary*. South Asia Books, Columbia, MO.

C. Cabezas, P. Resnik, and J. Stevens. 2001. Supervised sense tagging using Support Vector Machines. In *Proceedings of the Senseval-2 Workshop*, Toulouse, July.

T. Chklovski and R. Mihalcea. 2002. Building a sense tagged corpus with the Open Mind Word Expert. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia.

T. Pedersen. 2001. Machine learning with lexical features: The Duluth approach to Senseval-2. In *Proceedings of the Senseval-2 Workshop*, pages 139–142, Toulouse, July.

J. Raker and R. Shukla, editors. 1996. *Hippocrene Standard Dictionary English-Hindi Hindi-English (With Romanized Pronunciation)*. Hippocrene Books, New York, NY.

P. Resnik and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.