An Evaluation Exercise for Romanian Word Sense Disambiguation

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

Vivi Năstase
School of Computer Science
University of Ottawa
Ottawa, ON, Canada
vnastase@site.uottawa.ca

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

Doina Tătar
Department of Computer Science
Babeş-Bolyai University
Cluj-Napoca, Romania
dtatar@ubb.ro

Dan Tufiş
Romanian Academy Center
for Artificial Intelligence
Bucharest, Romania
tufis@racai.ro

Florentina Hristea
Department of Computer Science
University of Bucharest
Bucharest, Romania
fhristea@mailbox.ro

Abstract

This paper presents the task definition, resources, participating systems, and comparative results for a Romanian Word Sense Disambiguation task, which was organized as part of the SENSEVAL-3 evaluation exercise. Five teams with a total of seven systems were drawn to this task.

1 Introduction

SENSEVAL is an evaluation exercise of the latest word-sense disambiguation (WSD) systems. It serves as a forum that brings together researchers in WSD and domains that use WSD for various tasks. It allows researchers to discuss modifications that improve the performance of their systems, and analyze combinations that are optimal.

Since the first edition of the SENSEVAL competitions, a number of languages were added to the original set of tasks. Having the WSD task prepared for several languages provides the opportunity to test the generality of WSD systems, and to detect differences with respect to word senses in various languages.

This year we have proposed a Romanian WSD task. Five teams with a total of seven systems have tackled this task. We present in this paper the data used and how it was obtained, and the performance of the participating systems.

2 Open Mind Word Expert

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

To overcome the current lack of sense 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. Sense 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, who are asked to select the most appropriate sense for the target word in each sentence. The selection is made using checkboxes, which list all possible senses of the current target word, plus two additional choices, "unclear" and "none of the above." Although users are encouraged to select only one meaning per word, the selection of two or more senses is also possible. The results of the classification submitted by other users are not presented to avoid artificial biases.

3 Sense inventory

For the Romanian WSD task, we have chosen a set of words from three parts of speech - nouns, verbs and adjectives. Table 1 presents the number of words under each part of speech, and the average number of senses for each class.

The senses were (manually) extracted from a Romanian dictionary (Dicţionarul EXplicativ al limbii române - DEX (Coteanu et al., 1975)). These senses

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

Class	Number words	Avg senses (fine)	Avg senses (coarse)
Nouns	25	8.92	4.92
Verbs	9	8.7	4.6
Adjectives	5	9	4
Total	39	8.875	4.725

Table 1: Sense inventory

and their dictionary definitions were incorporated in the Open Mind Word Expert. For each annotation task, the contributors could choose from this list of 39 words. For each chosen word, the system displays the associated senses, together with their definitions, and a short (1-4 words) description of the sense. After the user gets familiarized with these senses, the system displays each example sentence, and the list of senses together with their short description, to facilitate the tagging process.

For the coarse grained WSD task, we had the option of using the grouping provided by the dictionary. A manual analysis however showed that some of the senses in the same group are quite distinguishable, while others that were separated were very similar.

For example, for the word *circulatie* (roughly, *circulation*). The following two senses are grouped in the dictionary:
2a. *movement, travel along a communication line/way*
2b. *movement of the sap in plants or the cytoplasm inside cells*
Sense 2a fits better with sense 1 of *circulation*:
1. *the event of moving about*
while sense 2b fits better with sense 3:
3. *movement or flow of a liquid, gas, etc. within a circuit or pipe*.

To obtain a better grouping, a linguist clustered the similar senses for each word in our list of forty. The average number of senses for each class is almost halved.

Notice that Romanian is a language that uses diacritics, and the the presence of diacritics may be crucial for distinguishing between words. For example *peste* without diacritics may mean *fish* or *over*. In choosing the list of words for the Romanian WSD task, we have tried to avoid such situations. Although some of the words in the list do have diacritics, omitting them does not introduce new ambiguities.

4 Corpus

Examples are extracted from the RoCo corpus, a 400 million words corpus consisting of a collection of Romanian newspapers collected on the Web over a three years period (1999-2002).

The corpus was tokenized and part-of-speech tagged using RACAI's tools (Tufis, 1999). The tokenizer recognizes and adequately segments various constructs: clitics, dates, abbreviations, multiword expressions, proper nouns, etc. The tagging followed the tiered tagging approach with the hidden layer of tagging being taken care of by Thorsten Brants' TNT (Brants, 2000). The upper level of the tiered tagger removed from the assigned tags all the attributes irrelevant for this WSD exercise. The estimated accuracy of the part-of-speech tagging is around 98%.

5 Sense Tagged Data

While several sense annotation schemes have been previously proposed, including single or dual annotations, or the "tag until two agree" scheme used during SENSEVAL-2, we decided to use a new scheme and collect four tags per item, which allowed us to conduct and compare inter-annotator agreement evaluations for two-, three-, and four-way agreement. The agreement rates are listed in Table 3. The two-way agreement is very high – above 90% – and these are the items that we used to build the annotated data set. Not surprisingly, four-way agreement is reached for a significantly smaller number of cases. While these items with four-way agreement were not explicitly used in the current evaluation, we believe that this represents a "platinum standard" data set with no precedent in the WSD research community, which may turn useful for a range of future experiments (for bootstrapping, in particular).

Agreement type	Total	(%)
TOTAL ITEMS	11,532	100%
At least two agree	10,890	94.43%
At least three agree	8,192	71.03%
At least four agree	4,812	41.72%

Table 3: Inter-agreement rates for two-, three-, and four-way agreement

Table 2 lists the target words selected for this task, together with their most common English translations. For each word, we also list the number of senses, as defined in the DEX sense inventory (collocations included), and the number of annotated examples made available to task participants.

Word	Main English translation	senses (fine)	senses (coarse)	Train size	Test size	Word	Main English translation	senses (fine)	senses (coarse)	Train size	Test size
NOUNS											
ac	needle	16	7	127	65	accent	accent	5	3	172	87
actiune	action	10	7	261	128	canal	channel	6	5	134	66
circuit	circuit	7	5	200	101	circulatie	circulation	9	3	221	114
coroana	crown	15	11	252	126	delfin	doplhin	5	4	31	15
demonstratie	demonstration	6	3	229	115	eruptie	eruption	2	2	54	27
geniu	genius	5	3	106	54	nucleu	nucleus	7	5	64	33
opozitie	opposition	12	7	266	134	perie	brush	5	3	46	24
pictura	painting	5	2	221	111	platforma	platform	11	8	226	116
port	port	7	3	219	108	problema	problem	6	4	262	131
proces	process	11	3	166	82	reactie	reaction	7	6	261	131
stil	style	14	4	199	101	timbru	stamp	7	3	231	116
tip	type	7	4	263	131	val	wave	15	9	242	121
valoare	value	23	9	251	125						
VERBS											
cistiga	win	5	4	227	115	citi	read	10	4	259	130
cobori	descend	11	6	252	128	conduce	drive	7	6	265	134
creste	grow	14	6	209	103	desena	draw	3	3	54	27
desface	untie	11	5	115	58	fierbe	boil	11	4	83	43
indulci	sweeten	7	4	19	10						
ADJECTIVES											
incet	slow	6	3	224	113	natural	natural	12	5	242	123
neted	smooth	7	3	34	17	oficial	official	5	3	185	96
simplu	simple	15	6	153	82						

Table 2: Target words in the SENSEVAL-3 Romanian Lexical Sample task

Team	System name	Reference (this volume)
Babes-Bolyai University, Cluj-Napoca (1)	ubb_nbc_ro	(Csomai, 2004)
Babes-Bolyai University, Cluj-Napoca (2)	UBB	(Serban and Tatar, 2004)
Swarthmore College	swat-romanian	(Wicentowski et al., 2004a)
Swarthmore College / Hong Kong Polytechnic University Hong Kong University of Science and Technology	swat-hk-romanian romanian-swat_hk-bo	(Wicentowski et al., 2004b)
University of Maryland, College Park	UMD_SST6	(Cabezas et al., 2004)
University of Minnesota, Duluth	Duluth-RomLex	(Pedersen, 2004)

Table 4: Teams participating in the SENSEVAL-3 Romanian Word Sense Disambiguation task

In addition to sense annotated examples, participants have been also provided with a large number of unlabeled examples. However, among all participating systems, only one system – described in (Serban and Tătar 2004) – attempted to integrate this additional unlabeled data set into the learning process.

6 Participating Systems

Five teams participated in this word sense disambiguation task. Table 4 lists the names of the participating systems, the corresponding institutions, and references to papers in this volume that provide detailed descriptions of the systems and additional analysis of their results.

There were no restrictions placed on the number of submissions each team could make. A total number of seven submissions was received for this task. Table 5 shows all the submissions for each team, and gives a brief description of their approaches.

7 Results and Discussion

Table 6 lists the results obtained by all participating systems, and the baseline obtained using the "most frequent sense" (MFS) heuristic. The table lists precision and recall figures for both fine grained and coarse grained scoring.

The performance of all systems is significantly higher than the baseline, with the best system performing at 72.7% (77.1%) for fine grained (coarse grained) scoring, which represents a 35% (38%) error reduction with respect to the baseline.

The best system (romanian-swat_hk-bo) relies on a Maximum Entropy classifier with boosting, using local context (neighboring words, lemmas, and their part of speech), as well as bag-of-words features for surrounding words.

Not surprisingly, several of the top performing systems are based on combinations of multiple sclassifiers, which shows once again that voting

System	Description
romanian-swat_hk-bo	Supervised learning using Maximum Entropy with boosting, using bag-of-words and n-grams around the head word as features
swat-hk-romanian	The swat-romanian and romanian-swat_hk-bo systems combined with majority voting.
Duluth-RLSS	An ensemble approach that takes a vote among three bagged decision trees, based on unigrams, bigrams and co-occurrence features
swat-romanian	Three classifiers: cosine similarity clustering, decision list, and Naive Bayes, using bag-of-words and n-grams around the head word as features combined with a majority voting scheme.
UMD_SST6	Supervised learning using Support Vector Machines, using contextual features.
ubb_nbc_ro	Supervised learning using a Naive Bayes learning scheme, and features extracted using a bag-of-words approach.
UBB	A k-NN memory-based learning approach, with bag-of-words features.

Table 5: Short description of the systems participating in the SENSEVAL-3 Romanian Word Sense Disambiguation task. All systems are supervised.

	Fine grained		Coarse grained	
System	P	R	P	R
romanian-swat_hk-bo	72.7%	72.7%	77.1%	77.1%
swat-hk-romanian	72.4%	72.4%	76.1%	76.1%
Duluth-RLSS	71.4%	71.4%	75.2%	75.2%
swat-romanian	71.0%	71.0%	74.9%	74.9%
UMD_SST6	70.7%	70.7%	74.6%	74.6%
ubb_nbc_ro	71.0%	68.2%	75.0%	72.0%
UBB	67.1%	67.1%	72.2%	72.2%
Baseline (MFS)	58.4%	58.4%	62.9%	62.9%

Table 6: System results on the Romanian Word Sense Disambiguation task.

schemes that combine several learning algorithms outperform the accuracy of individual classifiers.

8 Conclusion

A Romanian Word Sense Disambiguation task was organized as part of the SENSEVAL-3 evaluation exercise. In this paper, we presented the task definition, and resources involved, and shortly described the participating systems. The task drew the participation of five teams, and included seven different systems. The sense annotated data used in this exercise is available online from `http://www.senseval.org` and `http://teach-computers.org`.

Acknowledgments

Many thanks to all those who contributed to the Romanian Open Mind Word Expert project, making this task possible. Special thanks to Bogdan Harhata, from the Institute of Linguistics Cluj-Napoca, for building a coarse grained sense map. 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.

References

T. Brants. 2000. Tnt - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference, ANLP-2000*, Seattle, WA, May.

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

I. Coteanu, L. Seche, M. Seche, A. Burnei, E. Ciobanu, E. Contraş, Z. Creţa, V. Hristea, L. Mareş, E. Stîngaciu, Z. Ştefănescu, T. Ţugulea, I. Vulpescu, and T. Hristea. 1975. *Dicţionarul Explicativ al Limbii Române*. Editura Academiei Republicii Socialiste România.

D. Tufis. 1999. Tiered tagging and combined classifiers. In *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence*.