The MICRO-WNOP Corpus

October 26, 2006

This document describes the MICRO-WNOP corpus, which have been used in [Cerini et al., 2007].
This document is published on the Web at:
`http://www.unipv.it/wnop`
The MICRO-WNOP corpus is available for download at:
`http://www.unipv.it/wnop/micrownop.tgz`

1 The Micro-WNOp corpus

The MICRO-WNOP corpus is composed by 1105 WORDNET synsets. The corpus is divided into three parts:

- *Common* part. 110 synsets which the 5 evaluators have evaluated all together to align their evaluation criteria.
- *Group1* part. 496 synsets which have been evaluated by a group of three evaluators. Each evaluator has performed this part of the evaluation independently from the other ones.
- *Group2* part. 499 synsets which have been evaluated by the remaining two evaluators. Each evaluator has performed this part of the evaluation independently from the other one.

Two criteria have been adopted in the construction of the corpus:

- Opinion relevance: the corpus should contain enough synsets which are relevant respect to the opinion topic.
- WORDNET representativeness: the POS of the synsets in the corpus should be representative of the distribution of the synsets among the four POS.

To ensure the creation of a corpus composed by synsets which are relevant to the opinion topic, the *General Inquirer* (GI) lexicon [Stone et al., 1966] has been used to identify a set of terms which are relevant to the opinion topic.

The GI is a text analysis system that uses, in order to carry out its tasks, a lexicon with terms manually classified on a large number of categories, each one denoting the presence of a specific trait in a given term. The lexicon contains a total of 11788 terms, 1915 of them are labeled as Positive and 2291 are labeled as Negative (the remaining 7582 terms, not belonging either to Positive or Negative, can be considered to be (implicitly) labeled as Objective). The GI lexicon has been widely used in many works on the automatic determination of sentiment properties of terms [Kamps et al., 2004, Esuli and Sebastiani, 2005, Turney and Littman, 2003].

A list of 100 Positive terms have been created by randomly selecting Positive terms from the GI lexicon. The same process has been repeated for the Negative and Objective categories. The three terms lists obtained have been then converted into synsets lists by selecting from WordNet all the synsets where each term appears.

The synsets lists have been split into three parts to follow the partition of the Micro-WNOp corpus into *Common*, *Group1*, and *Group2*. Table 1 shows which terms have been selected and how they have been distributed among the parts of the corpus.

Micro-WNOp	GI Positive	GI Negative	GI Objective
Common	able, convince, famous, imperative, negotiate, real, sturdy, wonder, zest	abandon, boot, cynical, drab, guilt, killer, perplex, selfishness, thud, yearn, yelp	demand, flashlight, inflexible, monitor, semantic, thing
Group1	accommodate, adjust, affirmation, amiable, arisen, ball, blithe, careful, cleanliness, commendation, consistent, cure, desirable, earnestness, entertain, exuberance, flexible, gaiety, good, harmonious, humble, innovative, joke, luckily, meaningful, momentous, optimal, peaceful, precaution, profit, purposeful, religious, richness, satisfy, sincere, stately, sweet, treaty, upright, vitality, witty, woo, workmanship, worth, worth-while, zenith	accident, alarming, apathy, backwardness, betray, broke, catch, collide, confrontation, covet, deafness, demean, deter, disavowal, disproportionate, dungeon, evasion, fallout, flagrant, frown, harmful, hot, impetuous, inefficiency, interrupt, liar, manipulate, misuse, neutralize, ordeal, plod, punch, recoil, revert, sadness, shark, slanderous, spite, strife, suspend, trap, undignified, unqualified, vanity, weariness	abstract, aloud, arch, autumn, beyond, bubble, cat, chest, combustion, construe, crept, destruct, dollar, egypt, equation, export, forgot, generality, guardianship, hillside, immoderate, insurance, journal, ledger, made, medicare, museum, norm, origin, particular, phrase, pot, project, ratio, relation, rhythm, science, singly, spark, stoicism, supremacy, term, transom, unexpected, ventricle, wedding, workingmen
Group2	acquit, advance, alleviate, applause, augment, beneficent, brilliance, cheerful, collaborate, comprehensive, cozy, dedicate, distinctive, endorse, exact, fervor, foster, gifted, gratify, heroism, indispensable, interest, knowledge, majestic, mesh, nurture, palatial, polish, primarily, prosperity, refine, rescue, safety, sensitive, soundness, suit, thrive, understandable, vastness, wise, wonderful, workable, world-famous, worthiness, worthy	affliction, anguish, atrophy, bastard, blindness, butchery, circle, complicate, contradiction, criticize, defect, deride, difficulty, dishearten, disturb, enrage, expense, feign, forlorn, gloat, hell, ignorant, incompetence, injunction, irrational, loveless, misbehavior, murder, obsolete, overwhelming, presumptuous, rampant, reprehensible, rotten, scorn, shroud, smuggle, static, substitution, tempest, ugly, unguarded, untrustworthy, volatile, woe	advisor, ankle, assessment, bathe, both, cameroon, cell, client, conciliation, corn, dancer, discard, during, emphasize, exam, federal, fruit, government, hawaii, hundred, incontestability, investment, korea, liquor, marker, miner, nest, ocean, overwhelm, pension, plural, presumably, purple, recreate, researcher, runner, sheet, so, stadium, subsequent, tale, together, tying, untold, volunteer, why

Table 1: Repartition of GI terms on Micro-WNOp parts

Table 2 shows the repartition of synsets among POS in WordNet and in Micro-WNOp. The Micro-WNOp corpus has a smaller proportion of nouns respect to WordNet, this can be motivated by the large presence of proper nouns in WordNet, which instead are missing in the GI lexicon. The proportions among the other POS is respected, both in the whole corpus and in each of its three parts.

	Adjectives	Nouns	Verbs	Adverbs	Total
WordNet	18563 (16%)	79689 (69%)	3664 (3%)	13508 (12%)	115424
Whole Micro-WNOp	284 (26%)	500 (45%)	32 (3%)	289 (26%)	1105
Micro-WNOp *Common*	28 (25%)	51 (46%)	2 (2%)	29 (26%)	110
Micro-WNOp *Group1*	138 (28%)	214 (43%)	9 (2%)	135 (27%)	496
Micro-WNOp *Group2*	118 (24%)	235 (47%)	21 (4%)	125 (25%)	499

Table 2: Repartition of synsets among POS in WordNet and Micro-WNOp

2 Data format

The Micro-WNOpcorpus is available in a single tab-separated-values text file.

The file is made of three blocks, each one delimited by a pair of `# begin SectionName` and `# end SectionName` lines, where `SectionName` can be `Common`, `Group1`, or `Group2`.

A block consists of a list of evaluations, one line per synset. Every line contains one or more (three for Group1 and two for Group2) pair of values (*positivity* and *negativity* scores), followed by the list of terms, with POS and sense number, which belongs to the synsets in WordNet (the current release of Micro-WNOp refers to WordNet 2.0). The line of text immediatly following the `# begin SectionName` line is a comment line which describes in detail the meaning of each field in the block. The *objectivity* score for a synset can be computed as $1 - (positivity + negativity)$.

References

[Cerini et al., 2007] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. (2007). *Language resources and linguistic theory: Typology, second language acquisition, English linguistics (Forthcoming)*, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT.

[Esuli and Sebastiani, 2005] Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss analysis. In *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*, Bremen, DE.

[Kamps et al., 2004] Kamps, J., Marx, M., Mokken, R. o., and de Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. In *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, volume IV, pages 1115–1118, Lisbon, PT.

[Stone et al., 1966] Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press.

[Turney and Littman, 2003] Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.