Measuring Semantic Entropy

I. Dan Melamed
Dept. of Computer and Information Science
University of Pennsylvania
Philadelphia, PA, 19104, U.S.A.
`melamed@unagi.cis.upenn.edu`

Abstract

Semantic entropy is a measure of semantic ambiguity and uninformativeness. It is a graded lexical feature which may play a role anywhere lexical semantics plays a role. This paper presents a method for measuring semantic entropy using translational distributions of words in parallel text corpora. The measurement method is well-defined for all words, including function words, and even for punctuation.

1 Introduction

Semantic entropy is a measure of semantic ambiguity and uninformativeness. This paper presents a method for measuring semantic entropy using translational distributions of words in parallel text corpora. The measurement method is well-defined for all words, including function words, and even for punctuation. The hypothesis behind the measurement method is that semantically heavy words are more likely to have unique counterparts in other languages, so they tend to be translated more consistently than semantically lighter words. The consistency with which words are translated can be calculated from the translational distributions of words in parallel texts in two languages (bitexts). The translational distributions can be estimated using any reasonably good word translation model, such as those described in (BD+93; Che96) or in (Mel96b).

Semantic entropy is a graded lexical feature which may play a role anywhere lexical semantics plays a role. For example, semantic entropy can be interpreted as semantic ambiguity. On this interpretation, it can predict the difficulty of disambiguating the sense of a given word. Brown et al. (BD+91) present a word-sense disambiguation algorithm involving minimization of semantic entropy weighted by word frequency. Yarowsky (Yar93) compares the semantic entropy of homographs conditioned on different contexts. Another way to use semantic entropy for word-sense disambiguation is to allow disambiguation algorithms that favor precision over recall to ignore words with high semantic entropy. In the same vein, developers of interlinguas for machine translation can use semantic entropy to predict the required complexity of lexical elements of the representation.

Another interpretation of entropy is as the inverse of reliability. Machine learning algorithms may benefit from discounting the importance of data that has high entropy. For example, an algorithm learning selectional preferences may not want to generalize the statistical characteristics of "take into account" to other objects of "take," if it knows that "take" has high semantic entropy. I.e. the selectional preferences of "take" are hard to predict because it usually functions as a support verb. Resnik has used semantic entropy to explore selectional preferences, although he measured it in a different way (Res93).

Semantic entropy can help researchers decide not only how to work with words, but also which words to work with. Several applications in computational linguistics use stop-lists of unusual words. The canonical example is information retrieval systems, which routinely remove function words from queries. Another example is algorithms for mapping bitext correspondence at the word level. Such algorithms work better given a stop-list of words that are not likely to have cognates in other languages (Mel96a). For both of these applications, stop lists are typically constructed by rule of thumb and trial and error, uninformed by any theoretical underpinning. A common first approximation is the set of closed-class words. As will be illustrated in Section 3, semantic entropy may be a better indicator of function-wordhood than syntactic class.

The function/content word distinction also has a long history in psycholinguistics. For example, early research in the cognitive neuroscience of language suggested that function words and content words elicit qualitatively different event-related brain potentials (K&H83). Later work by the same researchers revealed that the differences were only quantitative and closely tied to word frequency

(Kim97). Section 4 explores the relationship between frequency and semantic entropy. It may be as useful or more useful to control semantic entropy in psycholinguistic experiments, the way that word frequency is usually controlled.

2 Method

2.1 Translational Distributions

The first step in measuring semantic entropy is to compute the translational distribution $\Pr(T|s)$ of each source word s in a bitext. A relatively simple method for estimating this distribution is described in (Mel96b). Briefly, the method works as follows:

1. Extract a set of aligned text segment pairs from a parallel corpus, e.g. using the techniques in (G&C91a) or in (Mel96a).

2. Construct an initial translation lexicon with likelihood scores attached to each entry, e.g. using the method in (Mel95) or in (G&C91).

3. Assume that words always translate one-to-one.

4. Armed with the current lexicon, greedily "link" each word token with its most likely translation in each pair of aligned segments.

5. Discard lexicon entries representing word pairs that are never linked.

6. Estimate the parameters of a maximum-likelihood word translation model.

7. Re-estimate the likelihood of each lexicon entry, using the number of times n its components co-occur, the number of times k that they are linked, and the probability $\Pr(k|n, model)$.

8. Repeat from Step 4 until the lexicon converges.

After the lexicon converges, Step 4 is repeated one last time, keeping track of how many times each English (source) word is linked to each French (target) word. Using the link frequencies $F(s,t)$ and the frequencies $F(s)$ of each English source word s, the maximum likelihood estimates of $\Pr(t|s)$, the probability that s translates to the French target word t, can be computed in the usual way: $\Pr(t|s) = F(s,t)/F(s)$.

2.2 Translational Entropy

The above method constructs translation lexicons containing only word-to-word correspondences. The best it can do for compound words like "au chaumage" and "right away" is to link their translation to the most representative part of the compound. For example, a typical translation lexicon may contain the entries "unemployed/chaumage" and "right/immédiatement." This behavior is quite suitable for our purposes, because we are interested only in the degree to which the translational probability mass is scattered over different target words,

not in the particular target words over which it is scattered.

The translational inconsistency of words can be computed following the principles of information theory[1]. In information theory, inconsistency is called entropy. Entropy is a functional of probability distribution functions (pdf's). If P is a pdf over the random variable X, then the entropy of P is defined as[2]

$$H(P) = -\sum_{x \in X} P(x) \log P(x).$$

Since probabilities are always between zero and one, their logarithms are always negative; the minus sign in the formula ensures that entropies are always positive.

The translational inconsistency of a source word s is proportional to the entropy $H(T|s)$ of its translational pdf $P(T|s)$:

$$H(T|s) = -\sum_{t \in T} P(t|s) \log P(t|s). \tag{1}$$

Note that $H(T|s)$ is *not* the same as the conditional entropy $H(T|S)$. The latter is a functional of the *entire* pdf of source words, whereas the former is a function of the *particular* source word s. The conditional entropy is actually a weighted sum of the individual translational entropies:

$$H(T|S) = \sum_{s \in S} P(s)H(T|s).$$

2.3 Null Links

All languages have words that don't translate easily into other languages, and paraphrases are common in translation. Most bitexts contain a number of word tokens in each text for which there is no obvious counterpart in the other text. Semantically light words are more likely to be paraphrased or translated non-literally. So, the frequency with which a particular word gets linked to nothing is an important factor in estimating its semantic entropy.

Ideally, a measure of translational inconsistency should be sensitive to which null links represent the same sense of a given source word and which ones represent different senses. Given that algorithms for making this distinction are currently beyond the state of the art, the simplest way to account for "null" links is to invent a special NULL word, and to pretend that all null links are actually links to NULL (BD+93). This heuristic produces undesired results, however, since it implies that the translation of a word which is never linked to anything is perfectly consistent. A better solution lies at the opposite extreme, in the assumption that each null link represents a *different* sense of the source word

---

[1] See (C&T91) for a good introduction.

[2] It is standard to use the shorthand notation $P(x)$ for $\Pr_P(X = x)$.

Table 1: *Parts of speech sorted by mean semantic entropy. Verbs include participles.*

Part of Speech (P)	number of types	$\bar{E}_P$	variance of E_P
prepositions	70	5.84	4.99
determiners	31	4.59	3.23
pronouns	37	4.14	2.86
conjunctions	11	2.77	1.40
punctuation	11	2.59	11.24
interjections	10	2.35	3.82
adverbs	972	2.21	2.36
verbs	7133	1.70	1.95
numerals	95	1.35	3.59
adjectives	5700	1.18	1.56
common nouns	10371	1.15	1.33
proper nouns	9280	0.34	0.53

Table 2: *Semantic entropy of punctuation has high variance.*

punctuation	frequency count	E
,	230810	12.27
–	8519	4.88
;	1922	2.73
-	105	2.36
:	763	1.91
(	11271	1.44
)	11264	1.42
..	70	1.38
!	1270	0.02
?	20231	0.00
.	278559	0.00

Table 3: *Adjectives sorted by semantic entropy.*

adjective	frequency count	E
other	9984	8.24
same	4913	8.04
such	5630	7.94
able	3217	7.39
few	2490	7.33
much	2402	7.22
certain	2109	7.22
least	1846	7.22
far	1760	7.04
free	3845	7.01
⋮	⋮	⋮
unemployed	319	5.50
corporate	475	5.50
hard	721	5.50
eastern	279	5.49
acting	282	5.49
now	286	5.48
coast	277	5.47
successful	448	5.47
late	588	5.47
strong	1161	5.42
⋮	⋮	⋮
reactionary	17	0.66
explanatory	17	0.66
psychiatric	17	0.66
biological	17	0.66
strategic	71	0.66
musical	10	0.64
intrinsic	10	0.64
august	10	0.64
cavalier	10	0.64
⋮	⋮	⋮

in question. Under this assumption, the contribution to the semantic entropy of s made by each null link is $-\frac{1}{F(s)}\log\frac{1}{F(s)}$. If $F(NULL|s)$ represents the number of times that s is linked to nothing, then the total contribution of all these null links to the semantic entropy of s is

$$N(s) = -F(NULL|s)\frac{1}{F(s)}\log\frac{1}{F(s)}$$
$$= P(NULL|s)\log F(s) \qquad (2)$$

The semantic entropy $E(s)$ of each word s accounts for both the null links and the non-null links of s:

$$E(s) = H(T|s) + N(s). \qquad (3)$$

3 Results

To estimate the semantic entropy of English words, roughly thirteen million words were used from the record of proceedings of the Canadian parliament ("Hansards"), which is available in English and in French. Before induction of the translation lexicon, both halves of the bitext were tagged for part of speech (POS) using Brill's transformation-based tagger (Bri92). The POS information was not used in the lexicon induction process but, after estimating the semantic entropies for all the English words in the corpus, the words were grouped into rough part-of-speech categories.

First, mean semantic entropy was compared across parts of speech. Table 1 lists the mean semantic entropies $\bar{E}_P$ for each part of speech P, sorted by $\bar{E}_P$, and the variance of each E_P. The table provides empirical evidence for the intuition that function words are translated less consistently than content words: The mean semantic entropy of each function-word POS is higher than that of any content-word POS. The table also shows that punctuation and interjections rank between the function words at the top and the content words at the bottom. This ranking is consistent with the intuition that punctuation and interjections have more semantic weight than function words, but less than content words.

43

Table 4: *Pronouns sorted by semantic entropy.*

pronoun	frequency count	E
's	11459	9.03
,	1636	7.28
there	24040	6.52
it	76036	6.43
themselves	1252	5.73
ourselves	615	5.60
what	20180	5.60
⋮	⋮	⋮
she	3281	3.32
me	5324	3.31
I	80901	3.24
him	4265	3.23
⋮	⋮	⋮
his	8	1.91
their	6	1.79
thee	4	1.39
thou	6	1.24

After analyzing the aggregated results, it was time to peek into the semantic entropy rankings within each POS. Several of these were particularly interesting. Table 2 explains the atypically high variance of the semantic entropy of punctuation.

End-of-sentence punctuation is used very consistently and almost identically in English and in French. So, the question mark, the exclamation mark and the period have almost no semantic entropy. In contrast, the two languages have different rules for comas and colons, especially around quotations. Comas and dashes are often used for similar purposes, so one is often translated as the other. Moreover, English comas are often lost in translation. For these reasons, the short Table 2 includes both the lowest and the highest semantic entropy values for English words in the Hansards.

Table 3 shows some of the adjectives, ranked by semantic Entropy. The top eight adjectives in the table say very little about the nouns that they might modify. They seem like thinly disguised function words that happen to appear in syntactic positions normally reserved for adjectives. Adjectives in the middle of the table are more typical, but they are less specific than the adjectives in the bottom third of the table.

Table 4 displays a sorted sample of the pronouns. Topping the list are the English possessive suffixes, which have no equivalent in French or in most other languages. Existential "there" is next. "It" is high on the list because of its frequent pleonastic function ("It is necessary to...."). These four pronouns are atypically functional. The most frequent of the thirty seven pronouns in the corpus, "I," is eleventh from the bottom of the list. The most consistently translated pronouns are the archaic forms "thee" and "thou."

Table 5: *Verbs with the highest semantic entropy.*

verb	participle?	frequency	E
do	-	37113	8.44
being	present	4166	7.75
going	present	1954	7.37
get	-	6888	7.14
be	-	245324	7.07
having	present	1989	7.02
made	past	4865	7.01
come	-	7088	6.99
concerned	past	2213	6.94
go	-	10079	6.87
involved	past	1784	6.77
making	present	1100	6.65
take	-	9249	6.59
put	-	4692	6.57
according	present	985	6.53
done	past	2580	6.52
doing	present	1192	6.49
taking	present	848	6.47
trying	present	837	6.44
stand	-	1939	6.44
given	past	2593	6.36
let	-	2975	6.36
given	past	2593	6.36
concerning	present	716	6.36
getting	present	649	6.36
dealing	present	870	6.30
saying	present	1262	6.28
may	-	5264	6.26
happen	-	3134	6.25
giving	present	653	6.23
make	-	11493	6.22
might	-	2755	6.20
told	past	663	6.11
taken	past	2061	6.10
clear	-	720	6.10
coming	present	899	6.09
become	-	2361	6.09
talking	present	526	6.08
directed	past	995	6.08
shall	-	796	6.08
brought	past	1159	6.03
bringing	present	515	6.03
putting	present	433	5.99
looking	present	480	5.99
been	past	3274	5.96
regarding	present	491	5.96
living	present	655	5.94
occur	-	674	5.92
agree	-	2770	5.89
bring	-	3075	5.84
fail	-	704	5.84
called	past	563	5.83
providing	present	489	5.81
using	present	477	5.81

The most interesting ranking of semantic entropies is among the verbs, including present and past participles. As shown in Table 5, verbs can have high entropies for several reasons. The verb with the highest semantic entropy by far is the functional verb place-holder "do." Very high on the list are various forms of the functional auxiliaries "be," "have," and "(be) going (to)," as well as the modals "may," "might," and "shall." The past participles "concerning," "involving," "according," "dealing," and "regarding" are near the top of the list because they occur most often as the heads of adjectival phrases modifying noun phrases, as in "the world according to NP", an English construction that is usually paraphrased in translation. "Try" and "let" are up there because they often serve as mere modal modifiers of a sentential argument. Most of the other verbs at the top of the list are light verbs. Verbs like "get," "make," "come," "take," "put," "stand," and "give" are often used as syntactic filler while most of the semantic content of the phrase is conveyed by their argument.

4 Discussion

The most in-depth study of semantic entropy and its applications to date was carried out by Resnik (Res93; Res95). Resnik's approach differs from the present one in three major ways. First, he defines semantic entropy over concepts, rather than over words. This definition is more useful for his particular applications, namely evaluating concept similarity and estimating selectional preferences. Second, in order to measure semantic similarity over concepts, his method requires a concept taxonomy, such as the Princeton WordNet (Mil90), which is grounded in the lexical ontology of a particular language. In contrast, the method presented in this paper requires a large bitext. Both kinds of resources are still available only for a limited number of languages, so only one of the two methods may be a viable option in any given situation. Third, Resnik's measure of information content is defined in terms of the logarithm of each concept's frequency in text, where the frequency of a concept is defined as the sum of the frequencies of words representing that concept in the taxonomy.

Given only monolingual data, log-frequency is a relatively good estimator of semantic entropy. Looking through the various tables in this paper, you may have noticed that words with higher entropy tend to have higher frequency. Semantic entropy, as measured here, actually correlates quite well with the logarithm of word frequency ($\rho = 0.79$). This correlation is to be expected, since the maximum possible entropy of a word with frequency f is $log(f)$, which is what Equation (3) evaluates to when a word is always linked to nothing. Yet the correlation is not perfect; simply sorting the words by frequency would

produce a suboptimal result. For instance, the most frequent pronoun in Table 4 is eleventh from the bottom of the list of thirty seven, because "I" has a very consistent meaning. Likewise, "going" has a higher entropy than "go" in Table 5, even though it is less than one fifth as frequent, because "going" can be used as a near-future tense marker whereas "go" has no such function. The best counter-example to the correlation between semantic entropy and log-frequency is the period, which is the most frequent token in the English Hansards and has a semantic entropy of zero.

The method presented here for measuring semantic entropy is sensitive to ontological and syntactic differences between languages. It is partly motivated by the observation that translators must paraphrase when the target language has no obvious equivalent for some word or syntactic construct in the source text. There are many more ways to paraphrase something than to translate it literally, and translators usually strive for variety in order to improve readability. That's why, for example, English light verbs have such high entropies even though there are many English verbs that are more frequent. The entropy of English light verbs would likely remain relatively high if English/Chinese bitexts were used instead of English/French, because the lexicalization patterns involving light verbs in English are particular to English. Reliance on this property of translated texts is a double-edged sword, however, due to the converse possibility that two languages share an unusual syntactic construct or an unusual bit of ontology. In that case, the relevant semantic entropies may be estimated too low. Ideally, semantic entropy should be estimated by averaging each source language of interest over several different target languages.

A more serious drawback of translational entropy as an estimate of semantic entropy is that words may be inconsistently translated either because they don't mean very much or because they mean several different things, or both. For example, WordNet 1.5 lists twenty six senses for the English verb "run." We would expect the different senses to have different translations in other languages, and we would expect several of these senses to occur in any sufficiently large bitext, resulting in a high estimate of semantic entropy for "run" (5.65 in the Hansards). Meanwhile, Table 5 shows that the English verb "be" is translated much less consistently[3] than "run," even though only nine senses are listed for it in WordNet. This is because "be" rarely conveys much information. It is useful to know about both of these components of semantic entropy, but it would be more useful to know about them separately (Ros97). This knowledge is contingent on knowledge of the elusive

---

[3]The semantic entropy metric is logarithmic. A difference of 1 represents a factor of 2.

$Pr(sense|word)$, which is currently the subject of much research (see, e.g. (N&L96) and references therein). Knowing $Pr(sense|word)$ would also improve Resnik's method, which so far has been forced to assume that this distribution is uniform (Res95).

5 Conclusion

The semantic entropy of a word can be interpreted as its semantic ambiguity and is inversely proportional to the word's information content, semantic weight, and consistency in translation. This paper presented an information-theoretic method for measuring the semantic entropy of any word in text, using translational distributions estimated from parallel text corpora. This measurement technique has produced entropy rankings that correspond well with intuitions about the relative semantic import of various words and word classes. The method can be implemented for any language for which a reasonably large bitext is available.

Acknowledgments

References

(Bri92) E. Brill, "A Simple Rule-Based Part of Speech Tagger," *3rd Conference on Applied Natural Language Processing*, pp. 152–155, 1992.

(BD+91) P. F. Brown, S. Della Pietra, V. Della Pietra, R. Mercer, "Word Sense Disambiguation using Statistical Methods", *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, Ca., 1991.

(BD+93) P. F. Brown, V. J. Della Pietra, S. A. Della Pietra & R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics 19*(2), 1993.

(Che96) S. Chen, *Building Probabilistic Models for Natural Language*, Ph.D. Thesis, Harvard University, May 1996.

(C&T91) T. M. Cover & J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, 1991.

(G&C91a) W. Gale, & K. W. Church, "A Program for Aligning Sentences in Bilingual Corpora" *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, Ca., 1991.

(G&C91) W. Gale & K. W. Church, "Identifying Word Correspondences in Parallel Texts," *DARPA SNL Workshop*, 1991.

(Kim97) A. Kim, personal communication, 1997.

(K&H83) M. Kutas & S. A. Hillyard, "Event-related brain potentials to grammatical errors and semantic anomalies," *Memory & Cognition 11*(5), 1983.

(Mel95) I. D. Melamed "Automatic Evaluation and Uniform Filter Cascades for Inducing N-best Translation Lexicons," *Third Workshop on Very Large Corpora*, Boston, MA, 1995.

(Mel96a) I. D. Melamed, "A Geometric Approach to Mapping Bitext Correspondence," *Conference on Empirical Methods in Natural Language Processing*, Philadelphia, U.S.A, 1996.

(Mel96b) I. D. Melamed, "Automatic Construction of Clean Broad-Coverage Translation Lexicons," *2nd Conference of the Association for Machine Translation in the Americas*, Montreal, Canada, 1996.

(Mil90) G. A. Miller (ed.), *WordNet: An On-Line Lexical Database. Special issue of International Journal of Lexicography 4*(3), 1990.

(N&L96) H. T. Ng & H. B. Lee, "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach," *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA, 1996.

(Res93) P. Resnik, *Selection and Information: A Class-Based Approach to Lexical Relationships*, PhD Thesis, University of Pennsylvania, Philadelphia, U.S.A, 1993.

(Res95) P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995.

(Ros97) J. Rosenzweig, personal communication, 1997.

(Yar95) D. Yarowsky, "One Sense Per Collocation," *DARPA Workshop on Human Language Technology*, Princeton, NJ, 1993.