Collocation Extraction beyond the Independence Assumption

Gerlof Bouma
Universität Potsdam, Department Linguistik
Campus Golm, Haus 24/35
Karl-Liebknecht-Straße 24–25
14476 Potsdam, Germany
gerlof.bouma@uni-potsdam.de

Abstract

In this paper we start to explore two-part collocation extraction association measures that do not estimate expected probabilities on the basis of the independence assumption. We propose two new measures based upon the well-known measures of mutual information and pointwise mutual information. Expected probabilities are derived from automatically trained Aggregate Markov Models. On three collocation gold standards, we find the new association measures vary in their effectiveness.

1 Introduction

Collocation extraction typically proceeds by scoring collocation candidates with an association measure, where high scores are taken to indicate likely collocationhood. Two well-known such measures are pointwise mutual information (PMI) and mutual information (MI). In terms of observing a combination of words w_1, w_2, these are:

$$i(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)\, p(w_2)}, \qquad (1)$$

$$I(w_1, w_2) = \sum_{\substack{x \in \{w_1, \neg w_1\} \\ y \in \{w_2, \neg w_2\}}} p(x, y)\, i(x, y). \qquad (2)$$

PMI (1) is the logged ratio of the observed bigramme probability and the expected bigramme probability under independence of the two words in the combination. MI (2) is the expected outcome of PMI, and measures how much information of the distribution of one word is contained in the distribution of the other. PMI was introduced into the collocation extraction field by Church and Hanks (1990). Dunning (1993) proposed the use of the likelihood-ratio test statistic, which is equivalent to MI up to a constant factor.

Two aspects of (P)MI are worth highlighting. First, the observed occurrence probability p_{obs} is compared to the expected occurrence probability p_{exp}. Secondly, the independence assumption underlies the estimation of p_{exp}.

The first aspect is motivated by the observation that interesting combinations are often those that are *unexpectedly* frequent. For instance, the bigramme *of the* is uninteresting from a collocation extraction perspective, although it probably is amongst the most frequent bigrammes for any English corpus. However, we can expect to frequently observe the combination by mere chance, simply because its parts are so frequent. Looking at p_{obs} and p_{exp} together allows us to recognize these cases (Manning and Schütze (1999) and Evert (2007) for more discussion).

The second aspect, the independence assumption in the estimation of p_{exp}, is more problematic, however, even in the context of collocation extraction. As Evert (2007, p42) notes, the assumption of "independence is extremely unrealistic," because it ignores "a variety of syntactic, semantic and lexical restrictions." Consider an estimate for $p_{\text{exp}}(\textit{the the})$. Under independence, this estimate will be high, as *the* itself is very frequent. However, with our knowledge of English syntax, we would say $p_{\text{exp}}(\textit{the the})$ is low. The independence assumption leads to overestimated expectation and *the the* will need to be very frequent for it to show up as a likely collocation. A less contrived example of how the independence assumption might mislead collocation extraction is when bigramme distribution is influenced by compositional, non-collocational, semantic dependencies. Investigating adjective-noun combinations in a corpus, we might find that *beige cloth* gets a high PMI, whereas *beige thought* does not. This does not make the former a collocation or multiword unit. Rather, what we would measure is the tendency to use colours with visible things and not with abstract objects. Syntactic and semantic

Proceedings of the ACL 2010 Conference Short Papers, pages 109–114,
Uppsala, Sweden, 11-16 July 2010.

associations between words are real dependencies, but they need not be *collocational* in nature. Because of the independence assumption, PMI and MI measure these syntactic and semantic associations just as much as they measure collocational association. In this paper, we therefore experimentally investigate the use of a more informed $p_{\exp}$ in the context of collocation extraction.

2 Aggregate Markov Models

To replace $p_{\exp}$ under independence, one might consider models with explicit linguistic information, such as a POS-tag bigramme model. This would for instance give us a more realistic $p_{\exp}(\textit{the the})$. However, lexical semantic information is harder to incorporate. We might not know exactly what factors are needed to estimate $p_{\exp}$ and even if we do, we might lack the resources to train the resulting models. The only thing we know about estimating $p_{\exp}$ is that we need more information than a unigramme model but less than a bigramme model (as this would make $p_{\text{obs}}/p_{\exp}$ uninformative). Therefore, we propose to use Aggregate Markov Models (Saul and Pereira, 1997; Hofmann and Puzicha, 1998; Rooth et al., 1999; Blitzer et al., 2005)[1] for the task of estimating $p_{\exp}$. In an AMM, bigramme probability is not directly modeled, but mediated by a hidden class variable c:

$$p_{\text{amm}}(w_2|w_1) = \sum_c p(c|w_1)p(w_2|c). \quad (3)$$

The number of classes in an AMM determines the amount of dependency that can be captured. In the case of just one class, AMM is equivalent to a unigramme model. AMMs become equivalent to the full bigramme model when the number of classes equals the size of the smallest of the vocabularies of the parts of the combination. Between these two extremes, AMMs can capture syntactic, lexical, semantic and even pragmatic dependencies.

AMMs can be trained with EM, using no more information than one would need for ML bigramme probability estimates. Specifications of the E- and M-steps can be found in any of the four papers cited above – here we follow Saul and Pereira (1997). At each iteration, the model components are updated according to:

$$p(c|w_1) \leftarrow \frac{\sum_w n(w_1,w)p(c|w_1,w)}{\sum_{w,c'} n(w_1,w)p(c'|w_1,w)}, \quad (4)$$

$$p(w_2|c) \leftarrow \frac{\sum_w n(w,w_2)p(c|w,w_2)}{\sum_{w,w'} n(w,w')p(c|w,w')}, \quad (5)$$

where $n(w_1,w_2)$ are bigramme counts and the posterior probability of a hidden category c is estimated by:

$$p(c|w_1,w_2) = \frac{p(c|w_1)p(w_2|c)}{\sum_{c'} p(c'|w_1)p(w_2|c')}. \quad (6)$$

Successive updates converge to a local maximum of the AMM's log-likelihood.

The definition of the counterparts to (P)MI without the independence assumption, the AMM-ratio and AMM-divergence, is now straightforward:

$$r_{\text{amm}}(w_1,w_2) = \log \frac{p(w_1,w_2)}{p(w_1)\,p_{\text{amm}}(w_2|w_1)}, \quad (7)$$

$$d_{\text{amm}}(w_1,w_2) = \sum_{\substack{x\in\{w_1,\neg w_1\}\\ y\in\{w_2,\neg w_2\}}} p(x,y)\, r_{\text{amm}}(x,y). \quad (8)$$

The free parameter in these association measures is the number of hidden classes in the AMM, that is, the amount of dependency between the bigramme parts used to estimate $p_{\exp}$. Note that AMM-ratio and AMM-divergence with one hidden class are equivalent to PMI and MI, respectively. It can be expected that in different corpora and for different types of collocation, different settings of this parameter are suitable.

3 Evaluation

3.1 Data and procedure

We apply AMM-ratio and AMM-divergence to three collocation gold standards. The effectiveness of association measures in collocation extraction is measured by ranking collocation candidates after the scores defined by the measures, and calculating average precision of these lists against the gold standard annotation. We consider the newly proposed AMM-based measures for a varying number of hidden categories. The new measures are compared against two baselines: ranking by frequency (p_{obs}) and random ordering. Because AMM-ratio and -divergence with one hidden class boil down to PMI and MI (and thus log-likelihood ratio), the evaluation contains an implicit comparison with

[1] These authors use very similar models, but with differing terminology and with different goals. The term AMM is used in the first and fourth paper. In the second paper, the models are referred to as Separable Mixture Models. Their use in collocation extraction is to our knowledge novel.

these canonical measures, too. However, the results will not be state-of-the-art: for the datasets investigated below, there are more effective extraction methods based on supervised machine learning (Pecina, 2008).

The first gold standard used is the German adjective-noun dataset (Evert, 2008). It contains 1212 A-N pairs taken from a German newspaper corpus. We consider three subtasks, depending on how strict we define true positives. We used the bigramme frequency data included in the resource. We assigned all types with a token count $\leq$5 to one type, resulting in AMM training data of 10k As, 20k Ns and 446k A-N pair types.

The second gold standard consists of 5102 German PP-verb combinations, also sampled from newspaper texts (Krenn, 2008). The data contains annotation for support verb constructions (FVGs) and figurative expressions. This resource also comes with its own frequency data. After frequency thresholding, AMMs are trained on 46k PPs, 7.6k Vs, and 890k PP-V pair types.

Third and last is the English verb-particle construction (VPC) gold standard (Baldwin, 2008), consisting of 3078 verb-particle pairs and annotation for transitive and intransitive idiomatic VPCs. We extract frequency data from the BNC, following the methods described in Baldwin (2005). This results in two slightly different datasets for the two types of VPC. For the intransitive VPCs, we train AMMs on 4.5k Vs, 35 particles, and 43k pair types. For the transitive VPCs, we have 5k Vs, 35 particles and 54k pair types.

All our EM runs start with randomly initialized model vectors. In Section 3.3 we discuss the impact of model variation due to this random factor.

3.2 Results

German A-N collocations The top slice in Table 1 shows results for the three subtasks of the A-N dataset. We see that using AMM-based $p_{\exp}$ initially improves average precision, for each task and for both the ratio and the divergence measure. At their maxima, the informed measures outperform both baselines as well as PMI and MI/log-likelihood ratio (# classes=1). The AMM-ratio performs best for 16-class AMMs, the optimum for AMM-divergence varies slightly.

It is likely that the drop in performance for the larger AMM-based measures is due to the AMMs learning the collocations themselves. That is, the AMMs become rich enough to not only capture the broadly applicative distributional influences of syntax and semantics, but also provide accurate $p_{\exp}$s for individual, distributionally deviant combinations – like collocations. An accurate $p_{\exp}$ results in a low association score.

One way of inspecting what kind of dependencies the AMMs pick up is to cluster the data with them. Following Blitzer et al. (2005), we take the 200 most frequent adjectives and assign them to the category that maximizes $p(c|w_1)$; likewise for nouns and $p(w_2|c)$. Four selected clusters (out of 16) are given in Table 2.[2] The esoteric class 1 contains ordinal numbers and nouns that one typically uses those with, including references to temporal concepts. Class 2 and 3 appear more semantically motivated, roughly containing human and collective denoting nouns, respectively. Class 4 shows a group of adjectives denoting colours and/or political affiliations and a less coherent set of nouns, although the noun cluster can be understood if we consider individual adjectives that are associated with this class. Our informal impression from looking at clusters is that this is a common situation: as a whole, a cluster cannot be easily characterized, although for subsets or individual pairs, one can get an intuition for why they are in the same class. Unfortunately, we also see that some actual collocations are clustered in class 4, such as *gelbe Karte* 'warning' (lit.: 'yellow card') and *dickes Auto* 'big (lit.: fat) car'.

German PP-Verb collocations The second slice in Table 1 shows that, for both subtypes of PP-V collocation, better $p_{\exp}$-estimates lead to *decreased* average precision. The most effective AMM-ratio and -distance measures are those equivalent to (P)MI. Apparently, the better $p_{\exp}$s are unfortunate for the extraction of the type of collocations in this dataset.

The poor performance of PMI on these data – clearly below frequency – has been noticed before by Krenn and Evert (2001). A possible explanation for the lack of improvement in the AMMs lies in the relatively high performing frequency baselines. The frequency baseline for FVGs is five times the

[2] An anonymous reviewer rightly warns against sketching an overly positive picture of the knowledge captured in the AMMs by only presenting a few clusters. However, the clustering performed here is only secondary to our main goal of improving collocation extraction. The model inspection should thus not be taken as an evaluation of the quality of the models as clustering models.

		# classes											
		1	2	4	8	16	32	64	128	256	512	Rnd	Frq
A-N													
category 1	r_{amm}	45.6	46.4	47.6	47.3	**48.3**	48.0	47.0	46.1	44.7	41.9	30.1	32.2
	d_{amm}	42.3	42.9	44.4	45.2	46.1	**46.5**	45.0	46.3	45.5	45.5		
category 1–2	r_{amm}	55.7	56.3	57.4	57.5	**58.1**	58.1	57.7	56.9	55.7	52.8	43.1	47.0
	d_{amm}	56.3	57.0	58.1	58.4	59.8	60.1	59.3	**60.6**	59.2	59.3		
category 1–3	r_{amm}	62.3	62.8	63.9	64.0	**64.4**	62.2	62.2	62.7	62.4	60.0	52.7	56.4
	d_{amm}	64.3	64.7	65.9	66.6	**66.7**	66.3	66.3	65.4	66.0	64.7		
PP-V													
figurative	r_{amm}	**7.5**	6.1	6.4	6.0	5.6	5.4	4.5	4.2	3.8	3.5	3.3	10.5
	d_{amm}	**14.4**	13.0	13.3	13.1	12.2	11.2	9.0	7.7	6.9	5.7		
FVG	r_{amm}	**4.1**	3.4	3.4	3.0	2.9	2.7	2.2	2.1	2.0	2.0	3.0	14.7
	d_{amm}	**15.3**	12.7	12.6	10.7	9.0	7.7	3.4	3.2	2.5	2.3		
VPC													
intransitive	r_{amm}	**9.3**	9.2	9.0	8.3	5.5	5.3					4.8	14.7
	d_{amm}	12.2	12.2	14.0	**16.3**	6.9	5.8						
transitive	r_{amm}	**16.4**	14.8	15.2	14.5	11.3	10.0					10.1	20.1
	d_{amm}	19.6	17.3	20.7	**23.8**	12.8	10.1						

Table 1: Average precision for AMM-based association measures and baselines on three datasets.

Cl	Adjective	Noun
1	*dritt* 'third', *erst* 'first', *fünft* 'fifth', *halb* 'half', *kommend* 'next', *laufend* 'current', *letzt* 'last', *nah* 'near', *paar* 'pair', *vergangen* 'last', *viert* 'fourth', *wenig* 'few', *zweit* 'second'	*Jahr* 'year', *Klasse* 'class', *Linie* 'line', *Mal* 'time', *Monat* 'month', *Platz* 'place', *Rang* 'grade', *Runde* 'round', *Saison* 'season', *Satz* 'sentence', *Schritt* 'step', *Sitzung* 'session', *Sonntag* 'Sunday', *Spiel* 'game', *Stunde* 'hour', *Tag* 'day', *Woche* 'week', *Wochenende* 'weekend'
2	*aktiv* 'active', *alt* 'old', *ausländisch* 'foreign', *betroffen* 'concerned', *jung* 'young', *lebend* 'alive', *meist* 'most', *unbekannt* 'unknown', *viel* 'many'	*Besucher* 'visitor', *Bürger* 'citizens', *Deutsche* 'German', *Frau* 'woman', *Gast* 'guest', *Jugendliche* 'youth', *Kind* 'child', *Leute* 'people', *Mädchen* 'girl', *Mann* 'man', *Mensch* 'human', *Mitglied* 'member'
3	*deutsch* 'German', *europäisch* 'European', *ganz* 'whole', *gesamt* 'whole', *international* 'international', *national* 'national', *örtlich* 'local', *ostdeutsch* 'East-German', *privat* 'private', *rein* 'pure', *sogenannt* 'so-called', *sonstig* 'other', *westlich* 'western'	*Betrieb* 'company', *Familie* 'family', *Firma* 'firm', *Gebiet* 'area', *Gesellschaft* 'society', *Land* 'country', *Mannschaft* 'team', *Markt* 'market', *Organisation* 'organisation', *Staat* 'state', *Stadtteil* 'city district', *System* 'system', *Team* 'team', *Unternehmen* 'enterprise', *Verein* 'club', *Welt* 'world'
4	*blau* 'blue', *dick* 'fat', *gelb* 'yellow', *grün* 'green', *linke* 'left', *recht* 'right', *rot* 'red', *schwarz* 'black', *white* 'weiß'	*Auge* 'eye', *Auto* 'car', *Haar* 'hair', *Hand* 'hand', *Karte* 'card', *Stimme* 'voice/vote'

Table 2: Selected adjective-noun clusters from a 16-class AMM.

random baseline, and MI does not outperform it by much. Since the AMMs provide a better fit for the more frequent pairs in the training data, they might end up providing too good p_{exp}-estimates for the true collocations from the beginning.

Further investigation is needed to find out whether this situation can be ameliorated and, if not, whether we can systematically identify for what kind of collocation extraction tasks using better p_{exp}s is simply not a good idea.

English Verb-Particle constructions The last gold standard is the English VPC dataset, shown in the bottom slice of Table 1. We have only used class-sizes up to 32, as there are only 35 particle types. We can clearly see the effect of the largest AMMs approaching the full bigramme model as average precision here approaches the random baseline. The VPC extraction task shows a difference between the two AMM-based measures: AMM-ratio does not improve at all, remaining below the frequency baseline. AMM-divergence, however, shows a slight decrease in precision first, but ends up performing above the frequency baseline for the 8-class AMMs in both subtasks.

Table 3 shows four clusters of verbs and particles. The large first cluster contains verbs that involve motion/displacement of the subject or object and associated particles, for instance *walk about* or *push away*. Interestingly, the description of the gold standard gives exactly such cases as negatives, since they constitute compositional verb-particle constructions (Baldwin, 2008). Classes 2 and 3 show syntactic dependencies, which helps

Cl	Verb	Particle
1	*break, bring, come, cut, drive, fall, get, go, lay, look, move, pass, push, put, run, sit, throw, turn, voice, walk*	*across, ahead, along, around, away, back, backward, down, forward, into, over, through, together*
2	*accord, add, apply, give, happen, lead, listen, offer, pay, present, refer, relate, return, rise, say, sell, send, speak, write*	*astray, to*
3	*know, talk, tell, think*	*about*
4	*accompany, achieve, affect, cause, create, follow, hit, increase, issue, mean, produce, replace, require, sign, support*	*by*

Table 3: Selected verb-particle clusters from an 8-class AMM on transitive data.

collocation extraction by decreasing the impact of verb-preposition associations that are due to PP-selecting verbs. Class 4 shows a third type of distributional generalization: the verbs in this class are all frequently used in the passive.

3.3 Variation due to local optima

We start each EM run with a random initialization of the model parameters. Since EM finds local rather than global optima, each run may lead to different AMMs, which in turn will affect AMM-based collocation extraction. To gain insight into this variation, we have trained 40 16-class AMMs on the A-N dataset. Table 4 gives five point summaries of the average precision of the resulting 40 'association measures'. Performance varies considerably, spanning 2–3 percentage points in each case. The models consistently outperform (P)MI in Table 1, though.

Several techniques might help to address this variation. One might try to find a good fixed way of initializing EM or to use EM variants that reduce the impact of the initial state (Smith and Eisner, 2004, a.o.), so that a run with the same data and the same number of classes will always learn (almost) the same model. On the assumption that an average over several runs will vary less than individual runs, we have also constructed a combined $p_{\exp}$ by averaging over 40 $p_{\exp}$s. The last column in Table 4 shows this combined estimator leads to good extraction results.

		Variation in avg precision					
		min	q1	med	q3	max	Comb
A-N							
cat 1	r_{amm}	46.5	47.3	47.9	48.4	49.1	48.4
	d_{amm}	44.4	45.4	45.8	46.1	47.1	46.4
cat 1–2	r_{amm}	56.7	57.2	57.9	58.2	59.0	58.2
	d_{amm}	58.1	58.8	59.2	59.4	60.4	60.0
cat 1–3	r_{amm}	63.0	63.7	64.2	64.6	65.3	64.6
	d_{amm}	65.2	66.0	66.4	66.6	67.6	66.9

Table 4: Variation on A-N data over 40 EM runs and result of combining p_{exp}s.

4 Conclusions

In this paper, we have started to explore collocation extraction beyond the assumption of independence. We have introduced two new association measures that do away with this assumption in the estimation of expected probabilities. The success of using these association measures varies. It remains to be investigated whether they can be improved more.

A possible obstacle in the adoption of AMMs in collocation extraction is that we have not provided any heuristic for setting the number of classes for the AMMs. We hope to be able to look into this question in future research. Luckily, for the AN and VPC data, the best models are not that large (in the order of 8–32 classes), which means that model fitting is fast enough to experiment with different settings. In general, considering these smaller models might suffice for tasks that have a fairly restricted definition of collocation candidate, like the tasks in our evaluation do. Because AMM fitting is unsupervised, selecting a class size is in this respect no different from selecting a suitable association measure from the canon of existing measures.

Future research into association measures that are not based on the independence assumption will also include considering different EM variants and other automatically learnable models besides the AMMs used in this paper. Finally, the idea of using an informed estimate of expected probability in an association measure need not be confined to (P)MI, as there are many other measures that employ expected probabilities.

Acknowledgements

This research was carried out in the context of the SFB 632 *Information Structure*, subproject D4: *Methoden zur interaktiven linguistischen Korpusanalyse von Informationsstruktur*.

References

Timothy Baldwin. 2005. The deep lexical acquisition of english verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

Timothy Baldwin. 2008. A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 1–2, Marrakech.

John Blitzer, Amir Globerson, and Fernando Pereira. 2005. Distributed latent variable models of lexical co-occurrences. In *Tenth International Workshop on Artificial Intelligence and Statistics*.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Stefan Evert. 2007. Corpora and collocations. Extended Manuscript of Chapter 58 of A. Lüdeling and M. Kytö, 2008, *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin.

Stefan Evert. 2008. A lexicographic evaluation of German adjective-noun collocations. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 3–6, Marrakech.

Thomas Hofmann and Jan Puzicha. 1998. Statistical models for co-occurrence data. Technical report, MIT. AI Memo 1625, CBCL Memo 159.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? a case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, Toulouse.

Brigitte Krenn. 2008. Description of evaluation resource – German PP-verb data. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 7–10, Marrakech.

Chris Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–57, Marrakech.

Mats Rooth, Stefan Riester, Detlef Prescher, Glenn Carrol, and Franz Beil. 1999. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD.

Lawrence Saul and Fernando Pereira. 1997. Aggregate and mixed-order markov models for statistical language processing. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 81–89.

Noah A. Smith and Jason Eisner. 2004. Annealing techniques for unsupervised statistical language learning. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.