Filtering and Measuring the Intrinsic Quality of Human Compositionality Judgments

Silvio Cordeiro[1,2], **Carlos Ramisch**[1], **Aline Villavicencio**[2]
[1] Aix Marseille Université, CNRS, LIF UMR 7279 (France)
[2] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
silvioricardoc@gmail.com carlos.ramisch@lif.univ-mrs.fr avillavicencio@inf.ufrgs.br

Abstract

This paper analyzes datasets with numerical scores that quantify the semantic compositionality of MWEs. We present the results of our analysis of crowdsourced compositionality judgments for noun compounds in three languages. Our goals are to look at the characteristics of the annotations in different languages; to examine intrinsic quality measures for such data; and to measure the impact of filters proposed in the literature on these measures. The cross-lingual results suggest that greater agreement is found for the extremes in the compositionality scale, and that outlier annotation removal is more effective than outlier annotator removal.

1 Introduction

Noun compounds (NCs) are a pervasive class of multiword expressions (MWEs) in many languages. They are conventionalized noun phrases whose semantics range from idiomatic to fully compositional interpretations (Nakov, 2013). In idiomatic NCs, the meaning of the whole does not come directly from the meaning of its parts (Baldwin and Kim, 2010). For instance, an *ivory tower* is not a physical place, but a non-realistic perspective. Its semantic interpretation has little or nothing to do with a literal *tower* built out of *ivory*.

The semantic compositionality of MWEs can be represented as a numerical score. Its value indicates how much individual words contribute to the meaning of the whole: e.g. *olive oil* may be seen as 80% *olive* and 100% *oil*, whereas *dead end* is 5% *dead* and 90% *end*. Low values imply idiomaticity, while high values imply compositionality. This information can be useful, e.g. to decide how an MWE should be translated (Cap et al., 2015).

Many datasets with compositionality judgments have been collected (e.g. Gurrutxaga and Alegria (2013) and McCarthy et al. (2003)). Reddy et al. (2011) asked Mechanical Turkers to annotate 90 English noun-noun compounds on a scale from 0 to 5 with respect to the literality of member words. This resource has been used to evaluate compositionality prediction systems (Salehi et al., 2015). A similar resource has been created for German by Roller et al. (2013), who propose two filtering techniques adopted in our experiments. Farahmand et al. (2015) created a dataset of 1042 compounds in English with binary annotations by 4 experts. The sum of the binary judgments has been used as a numerical score to evaluate compositionality prediction functions (Yazdani et al., 2015).

In this paper we report a cross-lingual examination of quality measures and filtering strategies for compound compositionality annotations. Using the dataset by Reddy et al. (2011) and its extension to English, French and Portuguese by Ramisch et al. (2016), we examine the filters reported by Roller et al. (2013) for German and assess whether they improve overall dataset quality in these three languages. This analysis aims at studying the distributions and characteristics of the human ratings, examining quality measures for the collected data, and measuring the impact of simple filtering techniques on these quality measures. In particular, we look at how the scores obtained are distributed across the compositionality scale, whether the scores of the individual components are correlated with

Proceedings of the 12th Workshop on Multiword Expressions, pages 32–37,
Berlin, Germany, August 7-12, 2016.

those of the compounds, and if there are cases of compounds that are more difficult to annotate than others. This paper is structured as follows: the three compositionality datasets are presented in §2. The quality measures and filtering strategies are described in §3 and the results of the analysis in §4. The paper concludes with discussion of the results and of future work (§5).

2 Compostionality Datasets

In this task, we built three datasets, in French (`fr`), Portuguese (`pt`) and English (`en`), containing human-annotated compositionality scores for 2-word NCs. Annotators were native speakers using an online non-timed questionnaire. They were shown a NC (e.g. `en` *ivory tower*) and three sentences where the compound occurs in a particular sense as context for disambiguation. They then provide three numerical scores in a scale from 0 (idiomatic) to 5 (compositional): the contribution of the head word to the whole ($s_{\mathbf{H}}$), the contribution of the modifier word to the whole ($s_{\mathbf{M}}$) and the contribution of both words to the whole ($s_{\mathbf{NC}}$). Each entry in the raw dataset can be represented as a tuple, containing:

- **annot**: identifier of a human annotator
- **H**: syntactic head of the NC (noun).
- **M**: syntactic modifier of the head, can be a noun (`en`) or an adjective (`en pt fr`).
- $s_{\mathbf{NC}}$: integer rating given by the human annotator **annot** assessing the compositionality of the NC.
- $s_{\mathbf{H}}$ and $s_{\mathbf{M}}$: Same as $s_{\mathbf{NC}}$ for the contribution of **H** and **M** to the meaning of the whole NC.
- **equiv**: A list of at least two paraphrases, synonyms or equivalent formulations. For instance, for *ivory tower*, common paraphrases include *privilege* and *utopia*.

The datasets contain comparable data collected using different methodologies due to the requirement and availability of native speakers. For `en` and `fr`, we used Amazon Mechanical Turk (AMT). Native `en` speakers abound on the platform, unlike for the other languages. For `fr`, the annotation took considerably longer, and the quality was not as good as `en`. For `pt`, not enough native speakers were found. Therefore, we developed a stand-alone interface for collecting `pt` judgments from volunteer annotators.

The `pt` and `fr` datasets contain 180 manually selected noun–adjective NCs each. The `en` dataset is the combination of 2 parts: REDDY (Reddy et al., 2011) with the original dataset downloaded from the authors' websites, and $\texttt{en}_{+}$, with 90 manually selected noun–noun and adjective–noun compounds.

For each NC, the final scores are calculated as the average of all its annotations. For instance, if the 5 annotations for the contribution of *ivory* to *ivory tower* were `[0,1,0,2,0]`, the final $\mu_{\mathbf{M}}$ score would be 3/5. In other words, we obtain 3 scores per compound (for the contribution of **H**, **M** and for both) by aggregating individual annotator's scores using the arithmetic mean μ.

3 Quality Measures and Filtering

To calculate the quality of a compositionality dataset, we adopt measures that reflect agreement among the different annotators. We also compare strategies for removing outlier data (which may have introduced noise among the judgments), and the impact of such removal in terms of data retention.

3.1 Quality Measures

Our hypothesis is that, if the task is well defined, native speaker annotators should *agree* with each other even in the absence of common training or expertise. Low agreement could be motivated by several reasons: unclear/vague instructions, ill-formed or highly polysemous NCs, etc.

Inter-Annotator Agreement (α) A classical measure of inter-annotator agreement is the kappa score, which not only considers the proportion of agreeing pairs but also factors out chance agreement. In our case, however, ratings are not categorical but ordinal, so the α score, would be more adequate (Artstein and Poesio, 2008). Nonetheless, it is only possible to calculate α when all annotators rate the same items, which is not our case. We do not report this score in our evaluation.

Standard Deviation (μ_σ and $P_{\sigma>1.5}$) The standard deviation σ of a score s estimates its

average distance from the mean. Therefore, if human annotators agree, σ should be low as they tend to provide similar ratings that converge toward the average score μ. On the other hand, high σ values indicate high disagreement. We propose two metrics:

- μ_σ Average standard deviation of a score s over all NCs.
- $P_{\sigma>1.5}$ Proportion of NCs in the dataset whose σ is higher than 1.5, following Reddy et al. (2011).

Rank Correlation (ρ_{oth}) If two annotators agree, the ranking of the NCs annotated by both must be similar. Since in an AMT like setting it is difficult to compare pairs of annotators because they may not annotate the same NCs, we compare the ranking of the NCs rated by an individual annotator a with the ranking of the same NCs according to the average of all other annotators $\mu_{\Omega-a}$. In order to consider only order differences rather than value differences, we use Spearman's rank correlation score, noted ρ_{oth}.

3.2 Filtering

This analysis focuses on the filtering strategies described by Roller et al. (2013).

Z-score Filtering Our first filtering strategy aims at removing outlier *annotations*, who perhaps were distracted or did not fully understand the meaning of a given NC. It is similar to the filter proposed by Roller et al. (2013). We remove individual NC annotations whose score s is more than z standard deviations σ away from the average $\mu_{\Omega-s}$ of other scores for the same compound. In other words, we remove a compound if $\frac{|s-\mu_{\Omega-s}|}{\sigma_{\Omega-s}} > z$ for one of the three ratings (NC, **H** or **M**).[1]

Spearman Filtering Our second filtering strategy aims at removing outlier *annotators*, e.g. spammers and non-native speakers. We define a threshold R on the rank-correlation with others ρ_{oth} below which we discard all scores provided by **annot**. This technique was also used by Roller et al. (2013).

[1]Differently from Roller et al. (2013), we do not include the score being filtered out in μ and σ estimates. Moreover, we apply the filter to the three scores of an NC simultaneously.

We employed two additional filters, not analyzed here. First, we only accept annotators who confirm they are native speakers by answering general demographic questions in an external form. Second, we manually remove annotators who provided malformed **equiv** answers, not only containing typos but also major errors, suggesting non-native status.

3.3 Filtering Impact

To determine the impact of outlier removal, we calculate two measures. The first one is used by Roller et al. (2013) in the context of data filtering. They consider the data retention rate DRR as the proportion of NCs in the dataset after filtering n_{filtered} with respect to the initial number of compounds n, that is, how much was retained after filtering. The second measure is the average number of annotations μ_n across all NCs.

4 Data Analysis

In this paper we discuss 4 questions in particular, related to the quality of the annotations.

Does filtering improve quality? Table 1 presents the quality results for all datasets, in their original form as well as filtered. The filter threshold configurations adopted in these analyses were, for `en` and `pt`: $z = 2.2$, $\rho = 0.5$, and for `fr`: $z = 2.5$, $\rho = 0.5$.

As can be seen in Table 1, filtering does improve the quality of the annotations. The more restrictive the filtering, the lower the number of annotations available, but also the higher is the agreement among annotators, for all languages. When no filtering is performed, there is an average of 14.92 annotations per compound, but average standard deviation values ranging from 1.08 to 1.21. The proportion of high standard deviation compounds is between 22.78% and 30.56%. With filtering, the number of annotations per compound drops to 13.03, but so does the average standard deviation, which becomes smaller than 1. The proportion of high standard deviation compounds is between 14% and 19%.

Figures 1 and 2 show the variation in the `pt` dataset's quality as a function of z-score and Spearman ρ choices, respectively. The former is quite effective at improving the quality of the annotations for these languages, while the

Dataset	μ_n	$\mu_{\sigma_{NC}}$	μ_{σ_H}	μ_{σ_M}	$P_{\sigma_{NC}>1.5}$	$P_{\sigma_H>1.5}$	$P_{\sigma_M>1.5}$	DRR
REDDY	15	0.99	0.94	0.89	5.56%	11.11%	8.89%	–
en$_+$ raw	18.8	1.17	1.05	1.18	18.89%	16.67%	27.78%	–
en$_+$ filter	15.7	0.87	0.66	0.88	3.33%	10.00%	14.44%	83.61%
fr raw	14.9	1.15	1.08	1.21	22.78%	24.44%	30.56%	–
fr filter	13	0.94	0.83	0.96	13.89%	15.00%	18.89%	87.34%
pt raw	31.8	1.22	1.09	1.20	14.44	17.22%	19.44%	–
pt filter	27.9	1.0	0.83	0.97	6.11%	8.89%	12.22%	87.81%

Table 1: Intrinsic quality measures for the raw and filtered datasets

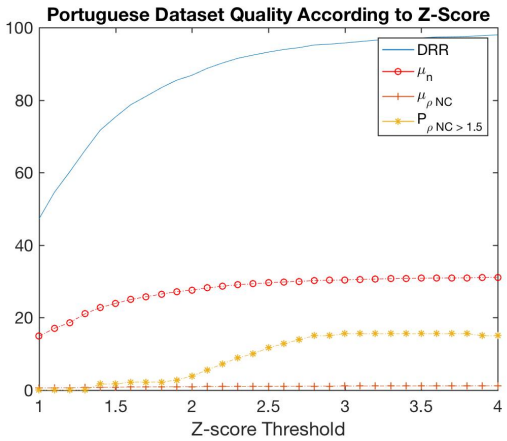

Figure 1: Quality of z-score filtering

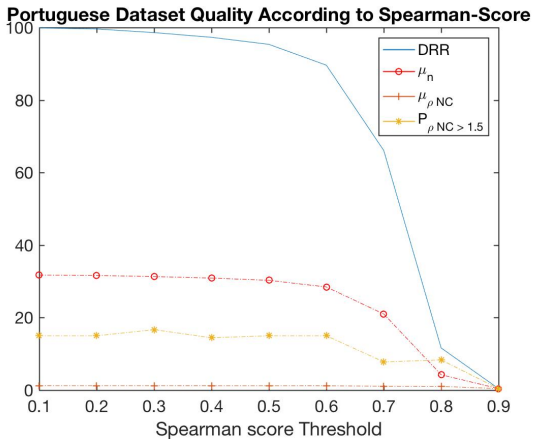

Figure 2: Quality of Spearman filtering

later does not seem to provide any real benefit. This differs from the results obtained by Roller et al. (2013) for German, but we see the same results consistently in our three datasets.

Are scores evenly distributed? Figure 3 shows the widespread distribution of compositionality scores of compounds (x-axis), compared with the combination of heads and modifiers (y-axis). This indicates that they are representative of the various compositionality scores, in a balanced manner.

Are the individual scores correlated? As can be seen in Figure 3, the average score for each compound can be reasonably approximated by the individual scores of head and modifier. Considering the goodness of fit measures R^2_{geom} and R^2_{arith} (for arithmetic and geometric means), we can see that the geometric model better represents the data. Whenever annotators judged an element of the compound as too idiomatic, they have also rated the whole compound as highly idiomatic.

Which NCs are harder to annotate? Figure 4 presents the standard deviation for each compound as a function of its average scores. One can visually attest that the least consensual compound judgments fall in the middle section of the graph. Even if we account for the fact that the extremities cannot follow a two-tailed distribution, those compounds still end up being easier than the ones in the middle.

5 Conclusions and Future Work

In this paper, we discussed the quality of human compositionality judgments, in English, French and Portuguese. We examined measures and filters for ensuring high agreement among annotators across languages. The cross-lingual results suggest that a greater agreement is obtained with outlier annotation removal than with outlier annotator removal, and that more agreement is found for the extremes of the compositionality scale.

Future work includes proposing a cross-lingual compositionality judgment protocol

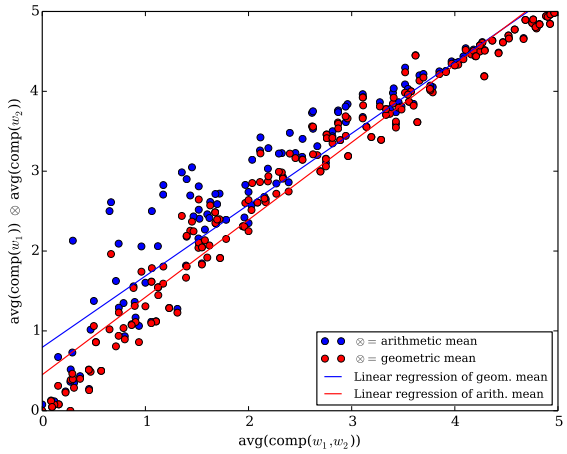

Figure 3: Distribution of $s_{\mathbf{H}} \otimes s_{\mathbf{M}}$ according to $s_{\mathbf{NC}}$ in `pt`.

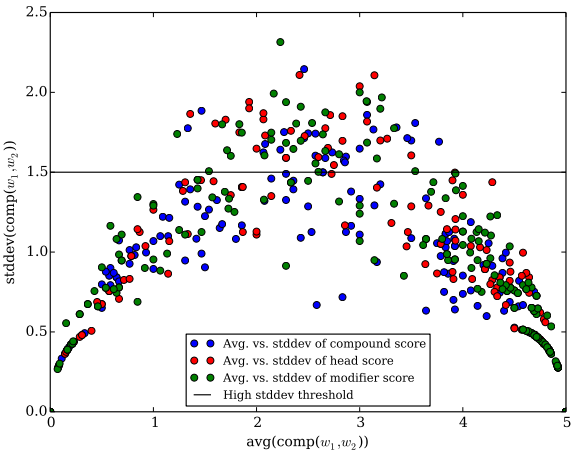

Figure 4: Distribution of $\sigma_{\mathbf{NC}}$ according to $\mu_{\mathbf{NC}}$ in `fr`.

that maximizes agreement among annotators. We also intend to examine the impact of factors like polysemy and concreteness of compound elements on annotator agreement. The complete resource, including filtered and raw data, is freely available.[2]

Acknowledgements

This work has been partly funded by projects "Simplificação Textual de Expressões Complexas", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91, PARSEME (Cost Action IC1207), PARSEME-FR (ANR-14-CERA-0001), AIM-WEST (FAPERGS-INRIA 1706-2551/13-7) and CNPq 482520/2012-4, 312114/2015-0.

[2] `http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/compounds`

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comp. Ling.*, 34(4):555–596.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.

Fabienne Cap, Manju Nirmal, Marion Weller, and Sabine Schulte im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proc. of the 11th Workshop on MWEs (MWE 2015)*, pages 19–28, Denver, Colorado, USA. ACL.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. In *Proc. of the 11th Workshop on MWEs (MWE 2015)*, pages 29–33, Denver, Colorado, USA. ACL.

Antton Gurrutxaga and Iñaki Alegria. 2013. Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *Proc. of the 9th Workshop on MWEs (MWE 2013)*, pages 116–125, Atlanta, GA, USA, Jun. ACL.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 73–80, Sapporo, Japan, Jul. ACL.

Preslav Nakov. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Nat. Lang. Eng. Special Issue on Noun Compounds*, 19(3):291–330.

Carlos Ramisch, Silvio Ricardo Cordeiro, Leonardo Zilio, Marco Idiart, Aline Villavicencio, and Rodrigo Wilkens. 2016. How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. In *Proc. of ACL 2016*. ACL. To appear.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011)*, November.

Stephen Roller, Sabine Schulte im Walde, and Silke Scheible. 2013. The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *Proc. of the 9th Workshop on MWEs (MWE 2013)*, pages 32–41, Atlanta, GA, USA, Jun. ACL.

Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015. The impact of multiword expression compositionality on machine translation evaluation. In *Proc. of the 11th Workshop on MWEs (MWE 2015)*, pages 54–59, Denver, Colorado, USA. ACL.

Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon, Portugal, September. Association for Computational Linguistics.