Sensicon: An Automatically Constructed Sensorial Lexicon

Serra Sinem Tekiroğlu
University of Trento
Fondazione Bruno Kessler
Trento, Italy
tekiroglu@fbk.eu

Gözde Özbal
Trento RISE
Trento, Italy
gozbalde@gmail.com

Carlo Strapparava
Fondazione Bruno Kessler
Trento, Italy
strappa@fbk.eu

Abstract

Connecting words with senses, namely, sight, hearing, taste, smell and touch, to comprehend the sensorial information in language is a straightforward task for humans by using commonsense knowledge. With this in mind, a lexicon associating words with senses would be crucial for the computational tasks aiming at interpretation of language. However, to the best of our knowledge, there is no systematic attempt in the literature to build such a resource. In this paper, we present a sensorial lexicon that associates English words with senses. To obtain this resource, we apply a computational method based on bootstrapping and corpus statistics. The quality of the resulting lexicon is evaluated with a gold standard created via crowdsourcing. The results show that a simple classifier relying on the lexicon outperforms two baselines on a sensory classification task, both at word and sentence level, and confirm the soundness of the proposed approach for the construction of the lexicon and the usefulness of the resource for computational applications.

1 Introduction

Sensorial information interpenetrates languages with various semantic roles in different levels since the main interaction instrument of humans with the outside world is the sensory organs. The transformation of the raw sensations that we receive through the sensory organs into our understanding of the world has been an important philosophical topic for centuries. According to a classification that dates back to Aristotle (Johansen, 1997), senses can be categorized into five modalities, namely, sight, hearing, taste, smell and touch. With the help of perception, we can process the data coming from our sensory receptors and become aware of our environment. While interpreting sensory data, we unconsciously use our existing knowledge and experience about the world to create a private experience (Bernstein, 2010).

Language has a significant role as our main communication device to convert our private experiences to shared representations of the environment that we perceive (Majid and Levinson, 2011). As a basic example, onomatopoeic words, such as *knock* or *woof*, are acquired by direct imitation of the sounds allowing us to share the experience of what we hear. As another example, where an imitation is not possible, is that giving a name to a color, such as *blue*, provides a tool to describe a visual feature of an object. In addition to the words that describe the direct sensorial features of objects, languages include many other lexical items that are connected to sensory modalities in various semantic roles. For instance, while some words can be used to describe a perception activity (e.g., *to sniff*, *to watch*, *to feel*), others can simply be physical phenomena that can be perceived by sensory receptors (e.g., *light*, *song*, *salt*, *smoke*).

Common usage of language, either written or spoken, can be very dense in terms of sensorial words. As an example, the sentence "*I felt the cold breeze.*" contains three sensorial words: *to feel* as a perception activity, *cold* as a perceived sensorial feature and *breeze* as a physical phenomenon. The connection to the sense modalities of the words might not be mutually exclusive, that is to say a word can be associated with more than one senses. For instance, the adjective *sweet* could be associated with both the senses of *taste* and *smell*. While we, as humans, have the ability to connect words with senses intuitively by using our commonsense knowledge, it is not straightforward for machines to interpret sensorial information.

Making use of a lexicon containing sensorial words could be beneficial for many computational scenarios. Rodriguez-Esteban and Rzhetsky

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1511–1521,
October 25-29, 2014, Doha, Qatar.

(2008) report that using words related to senses in a text could clarify the meaning of an abstract concept by facilitating a more concrete imagination. To this respect, an existing text could be automatically modified with sensory words for various purposes such as attracting attention or biasing the audience towards a specific concept. Additionally, sensory words can be utilized to affect private psychology by inducing a positive or negative sentiment (Majid and Levinson, 2011). For instance, de Araujo et al. (2005) show that the pleasantness level of the same odor can be altered by labeling it as *body odor* or *cheddar cheese*. As another motivation, the readability and understandability of text could also be enhanced by using sensory words (Rodriguez-Esteban and Rzhetsky, 2008). A compelling use case of a sensorial lexicon is that automatic text modification to change the density of a specific sense could help people with sensory disabilities. For instance, while teaching a concept to a congenitally blind child, an application that eliminates color-related descriptions would be beneficial. A sensorial lexicon could also be exploited by search engines to personalize the results according to user needs.

Advertising is another broad area which would benefit from such a resource especially by using synaesthesia[1], as it strengthens creative thinking and it is commonly exploited as an imagination boosting tool in advertisement slogans (Pricken, 2008). As an example, we can consider the slogans "*The taste of a paradise*" where the sense of sight is combined with the sense of taste or "*Hear the big picture*" where *sight* and *hearing* are merged.

Various studies have been conducted both in computational linguistics and cognitive science that build resources associating words with several cognitive features such as abstractness-concreteness (Coltheart, 1981; Turney et al., 2011), emotions (Strapparava and Valitutti, 2004; Mohammad and Turney, 2010), colors (Özbal et al., 2011; Mohammad, 2011) and imageability (Coltheart, 1981). However, to the best of our knowledge, there is no attempt in the literature to build a resource that associates words with senses. In this paper, we propose a computational method to automatically generate a sensorial lexicon that associates words in English with senses. Our method consists of two main steps. First, we generate a set of seed words for each sense category with the help of a bootstrapping approach. In the second step, we exploit a corpus based probabilistic technique to create the final lexicon. We evaluate this lexicon with the help of a gold standard that we obtain by using the crowdsourcing service of CrowdFlower[2].

The sensorial lexicon, which we named Sensicon, embodies 22,684 English lemmas together with their part-of-speech (POS) information that have been linked to one or more of the five senses. Each entry in this lexicon consists of a lemma-POS pair and a score for each sensory modality that indicates the degree of association. For instance, the verb *stink* has the highest score for *smell* as expected while the scores for the other four senses are very low. The noun *tree*, which is a concrete object and might be perceived by multiple senses, has high scores for *sight*, *touch* and *smell*.

The rest of the paper is organized as follows. We first review previous work relevant to this task in Section 2. Then in Section 3, we describe the proposed approach in detail. In Section 4, we explain the annotation process that we conducted and the evaluation strategy that we employed. Finally, in Section 5, we draw our conclusions and outline possible future directions.

2 Related Work

Since to the best of our knowledge there is no attempt in the literature to automatically associate words with human senses, in this section we will summarize the most relevant studies that focused on linking words with various other cognitive features.

There are several studies focusing on word-emotion associations. WordNet Affect Lexicon (Strapparava and Valitutti, 2004) maps WordNet (Fellbaum, 1998) synsets to various cognitive features (e.g., emotion, mood, behaviour). This resource is created by using a small set of synsets as seeds and expanding them with the help of semantic and lexical relations among these synsets. Yang et al. (2007) propose a collocation model with emoticons instead of seed words while creating an emotion lexicon from a corpus. Perrie et al. (2013) build a word-emotion association lexicon by using subsets of a human-annotated lexicon as seed sets. The authors use frequencies, counts, or unique seed words extracted from an n-gram corpus to create lexicons in different sizes. They pro-

[1] American Heritage Dictionary (`http://ahdictionary.com/`) defines synaesthesia in linguistics as the description of one kind of sense impression by using words that normally describe another.

[2] `http://www.crowdflower.com/`

pose that larger lexicons with less accurate generation method perform better than the smaller human annotated lexicons. While a major drawback of manually generated lexicons is that they require a great deal of human labor, crowdsourcing services provide an easier procedure for manual annotations. Mohammad and Turney (2010) generate an emotion lexicon by using the crowdsourcing service provided by Amazon Mechanical Turk[3] and it covers 14,200 term-emotion associations.

Regarding the sentiment orientations and subjectivity levels of words, Sentiwordnet (Esuli and Sebastiani, 2006) is constructed as an extension to WordNet and it provides sentiments in synset level. Positive, negative and neutral values are assigned to synsets by using ternary classifiers and synset glosses. Another study that has been inspirational for the design of our approach is Banea et al. (2008). The authors generate a subjectivity lexicon starting with a set of seed words and then using a similarity measure among the seeds and the candidate words.

Another cognitive feature relevant to sensorial load of the words is the association between colors and words. Mohammad (2011) builds a color-word association lexicon by organizing a crowdsourcing task on Amazon Mechanical Turk. Instead, Özbal et al. (2011) aim to automate this process and propose three computational methods based on image analysis, language models and latent semantic analysis (LSA) (Landauer and Dumais, 1997). The authors compare these methods against a gold standard obtained by the crowdsourcing service of Amazon Mechanical Turk. The best performance is obtained by using image features while LSA performs slightly better than the baseline.

Finally, there have been efforts in the literature about the association of words with their abstractness-concreteness and imageability levels. MRC Psycholinguistic Database (Coltheart, 1981) includes abstractness-concreteness and imageability ratings of a small set of words determined according to psycholinguistic experiments. Turney et al. (2011) propose to use LSA similarities of words with a set of seed words to automatically calculate the abstractness and concreteness degrees of words.

[3] http://www.mturk.com/mturk

3 Automatic Association of Senses with Words

We adopt a two phased computational approach to construct a large sensorial lexicon. First, we employ a bootstrapping strategy to generate a sufficient number of sensory seed words from a small set of manually selected seed words. In the second phase, we perform a corpus based probabilistic method to estimate the association scores to build a larger lexicon.

3.1 Selecting Seed Words

The first phase of the lexicon construction process aims to collect *sensorial seed words*, which are directly related to senses (e.g., *sound*, *tasty* and *sightedness*). To achieve that, we utilized a lexical database called FrameNet (Baker et al., 1998), which is built upon *semantic frame*s of concepts in English and lexical units (i.e., words) that evoke these frames. The basic idea behind this resource is that meanings of words can be understood on the basis of a semantic frame. A semantic frame consists of semantic roles called frame elements, which are manually annotated in more than 170,000 sentences. We have considered FrameNet to be especially suitable for the collection of sensorial seed words since it includes semantic roles and syntactic features of sensational and perceptional concepts.

In order to determine the seed lemma-POS pairs in FrameNet, we first manually determined 31 frames that we found to be highly connected to senses such as *Hear*, *Color*, *Temperature* and *Perception_experience*. Then, we conducted an annotation task and asked 3 annotators to determine which senses the lemma-POS pairs evoking the collected frames are associated with. At the end of this task, we collected all the pairs (i.e. 277) with 100% agreement to constitute our initial seed set. This set contains 277 lemma-POS pairs associated with a specific sense such as the verb *click* with *hearing*, the noun *glitter* with *sight* and *aromatic* with *smell*.

3.2 Seed Expansion via Bootstrapping

In this step, we aim to extend the seed list that we obtained from FrameNet with the help of a bootstrapping approach. To achieve that, we adopt a similar approach to Dias et al. (2014), who propose a repetitive semantic expansion model to automatically build temporal associations of synsets in WordNet. Figure 1 provides an overview of the bootstrapping process. At each iteration, we

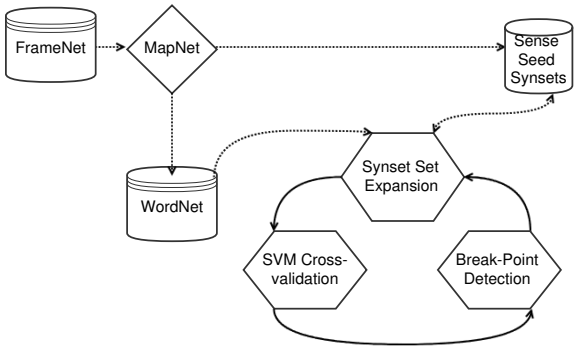

Figure 1: Bootstrapping procedure to expand the seed list.

first expand the seed list by using semantic relations provided by WordNet. We then evaluate the accuracy of the new seed list for sense classification by means of cross-validation against WordNet glosses. For each sense, we continue iterating until the cross-validation accuracy becomes stable or starts to decrease. The following sections explain the whole process in detail.

3.2.1 Extending the Seed List with WordNet

While the initial sensory seed list obtained from FrameNet contains only 277 lemma-POS pairs, we extend this list by utilizing the semantic relations provided by WordNet. To achieve that, we first map each lemma-POS pair in the seed list to WordNet synsets with the help of MapNet (Tonelli and Pighin, 2009), which is a resource providing direct mapping between WordNet synsets and FrameNet lexical units. Then, we add to the list the synsets that have WordNet relations *direct antonymy*, *similarity*, *derived-from*, *derivationally-related*, *pertains-to*, *attribute* and *also-see* with the already existing seeds. For instance, we add the synset containing the verb *laugh* for the synset of the verb *cry* with the relation *direct antonymy*, or the synset containing the adjective *chilly* for the synset of the adjective *cold* with the relation *similarity*. We prefer to use these relations as they might allow us to preserve the semantic information as much as possible during the extension process. It is worth mentioning that these relations were also found to be appropriate for preserving the affective connotation by Valitutti et al. (2004). Additionally, we use the relations *hyponym* and *hyponym-instance* to enrich the seed set with semantically more specific synsets. For instance, for the noun seed *smell*, we expand the list with the hyponyms of its synset such as the nouns *bouquet*, *fragrance*, *fragrancy*, *redolence* and *sweetness*.

3.2.2 Cross-validation of Sensorial Model

After obtaining new synsets with the help of WordNet relations in each bootstrapping cycle, we build a five-class sense classifier over the seed synsets defined by their glosses provided in WordNet. Similarly to Dias et al. (2014), we assume that the sense information of sensorial synsets is preserved in their definitions. Accordingly, we employ a support vector machine (SVM) (Boser et al., 1992; Vapnik, 1998) model with second degree polynomial kernel by representing the gloss of each synset as a vector of lemmas weighted by their counts. For each synset, its gloss is lemmatized by using Stanford Core NLP[4] and cleaned from the stop words. After each iteration cycle, we perform a 10-fold cross-validation in the updated seed list to detect the accuracy of the new sensorial model. For each sense class, we continue iterating and thereby expanding the seed list until the classifier accuracy steadily drops.

Table 1 lists the *precision* (*P*), *recall* (*R*) and *F1* values obtained for each sense after each iteration until the bootstrapping mechanism stops. While the iteration number is provided in the first column, the values under the last column group present the micro-average of the resulting multi-class classifier. The change in the performance values of each class in each iteration reveals that the number of iterations required to obtain the seed lists varies for each sense. For instance, the F1 value of *touch* continues to increase until the fourth cycle whereas *hearing* records a sharp decrease after the first iteration.

After the bootstrapping process, we create the final lexicon by repeating the expansion for each class until the optimal number of iterations is reached. The last row of Table 1, labeled as *Final*, demonstrates the accuracy of the classifier trained and tested on the final lexicon, i.e., using the seeds selected after iteration 2 for *Sight*, iteration 1 for *Hearing*, iteration 3 for *Taste* and *Smell* and iteration 4 for *Touch*. According to F1 measurements of each iteration, while *hearing* and *taste* have a lower value for the final model, *sight*, *smell* and *touch* have higher results. It should also be noted that the micro-average of the F1 values of the final model shows an increase when compared to the third iteration, which has the highest average F1 value among the iterations. At the end

[4] http://nlp.stanford.edu/software/corenlp.shtml

of this step we have a seed synset list consisting of 2572 synsets yielding the highest performance when used to learn a sensorial model.

3.3 Sensorial Lexicon Construction Using Corpus Statistics

After generating the seed lists consisting of synsets for each sense category with the help of a set of WordNet relations and a bootstrapping process, we use corpus statistics to create our final sensorial lexicon. More specifically, we exploit a probabilistic approach based on the co-occurrence of the seeds and the candidate lexical entries. Since working on the synset level would raise the data sparsity problem in synset tagged corpora such as SemCor (Miller et al., 1993) and we need a corpus that provides sufficient statistical information, we migrate from synset level to lexical level. Accordingly, we treat each POS role of the same lemmas as a distinct seed and extract 4287 lemma-POS pairs from 2572 synsets. In this section, we explain the steps to construct our final sensorial lexicon in detail.

3.3.1 Corpus and Candidate Words

As a corpus, we use a subset of English GigaWord 5th Edition released by Linguistic Data Consortium (LDC)[5]. This resource is a collection of almost 10 million English newswire documents collected in recent years, whose content sums up to nearly 5 billion words. The richly annotated GigaWord data comprises automatic parses obtained with the Stanford parser (Klein and Manning, 2003) so that we easily have access to the lemma and POS information of each word in the resource. For the scope of this study, we work on a randomly chosen subset that contains 79800 sentences and we define a co-occurrence event as the co-existence of a candidate word and a seed word within a window of 9 words(the candidate word, 4 words to its left and 4 words to its right). In this manner, we analyze the co-occurrence of each unique lemma-POS pair in the corpus with the sense seeds. We eliminate the candidates which have less than 5 co-occurrences with the sense categories.

3.3.2 Normalized Pointwise Mutual Information

For the co-occurrence analysis of the candidate words and seeds, we use pointwise mutual information (PMI), which is simply a measure of association between the probability of the co-occurrence of two events and their individual probabilities when they are assumed to be independent (Church and Hanks, 1990). PMI can be exploited as a semantic similarity measure (Han et al., 2013) and it is calculated as:

$$PMI(x,y) = \log\left[\frac{p(x,y)}{p(x)p(y)}\right] \quad (1)$$

To calculate the PMI value of a candidate word and a specific sense, we consider *p(x)* as the probability of the candidate word to occur in the corpus. Therefore, *p(x)* is calculated as $p(x) = c(x)/N$, where c(x) is the total count of the occurrences of the candidate word *x* in the corpus and N is the total co-occurrence count of all words in the corpus. Similarly, we calculate *p(y)* as the total occurrence count of all the seeds for the sense considered (*y*). *p(y)* can thus be formulated as $c(y)/N$. *p(x,y)* is the probability of the co-occurrence of a candidate word *x* with a sense event *y*.

A major shortcoming of PMI is its sensitivity for low frequency data (Bouma, 2009). As one possible solution, the author introduces Normalized Pointwise Mutual Information (NPMI), which normalizes the PMI values to the range (-1, +1) with the following formula:

$$NPMI(x,y) = \frac{PMI(x,y)}{-\log p(x,y)} \quad (2)$$

We adopt the proposed solution and calculate NPMI values for each candidate word and five sense events in the corpus. Sensicon covers 22,684 lemma-POS pairs and a score for each sense class that denotes their association degrees.

4 Evaluation

To evaluate the performance of the sensorial classification and the quality of Sensicon, we first created a gold standard with the help of a crowdsourcing task. Then, we compared the decisions coming from Sensicon against the gold standard. In this section, we explain the annotation process that we conducted and the evaluation technique that we adopted in detail. We also provide a brief discussion about the obtained results.

4.1 Crowdsourcing to Build a Gold Standard

The evaluation phase of Sensicon requires a gold standard data to be able to conduct a meaningful assessment. Since to our best knowledge there is no resource with sensory associations of words or

[5] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011T07

	Sight			Hearing			Taste			Smell			Touch			Micro-average		
It#	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	.873	.506	.640	.893	.607	**.723**	.716	.983	.828	.900	.273	.419	.759	.320	.451	.780	.754	.729
2	.666	.890	**.762**	.829	.414	.552	.869	.929	.898	.746	.473	.579	.714	.439	.543	.791	.787	.772
3	.643	.878	.742	.863	.390	.538	.891	.909	**.900**	.667	.525	**.588**	.720	.482	.578	.796	.786	**.776**
4	.641	.869	.738	.832	.400	.540	.866	.888	.877	.704	.500	.585	.736	.477	**.579**	.784	.774	.765
5	.640	.869	.737	.832	.400	.540	.866	.888	.877	.704	.500	.585	.738	.474	.578	.784	.774	.764
Final	.805	.827	.816	.840	.408	.549	.814	.942	.873	.685	.534	.600	.760	.582	.659	.800	.802	.790

Table 1: Bootstrapping cycles with validation results.

majority class	3	4	5	6	7	8	9	10
word	0	0.98	3.84	9.96	11.63	16.66	34.41	12.42
sentence	0.58	2.35	7.07	10.91	13.27	15.63	21.23	16.51

Table 2: Percentage of words and sentences in each majority class.

sentences, we designed our own annotation task using the crowdsourcing service of CrowdFlower. For the annotation task, we first compiled a collection of sentences to be annotated. Then, we designed two questions that the annotators were expected to answer for a given sentence. While the first question is related to the sense association of a whole sentence, the second asks the annotators to collect a fine-grained gold standard for word-sense associations.

We collected a dataset of 340 sentences consisting of 300 advertisement slogans from 11 advertisement categories (e.g., fashion, food, electronics) and 40 story sentences from a story corpus. We collected the slogans from various online resources such as `http://slogans.wikia.com/wiki` and `http://www.adslogans.co.uk/`. The story corpus is generated as part of a dissertation research (Alm, 2008) and it provides stories as a collection of sentences.

In both resources, we first determined the candidate sentences that had at least five tokens and contained at least one adjective, verb or noun. In addition, we replaced the brand names in the advertisement slogans with *X* to prevent any bias. For instance, the name of a well-known restaurant in a slogan might cause a bias towards *taste*. Finally, the slogans used in the annotation task were chosen randomly among the candidate sentences by considering a balanced number of slogans from each category. Similarly, 40 story sentences were selected randomly among the candidate story sentences. To give a more concrete idea, for our dataset we obtained an advertisement slogan such as "*X's Sugar Frosted Flakes They're Great!*" or a story sentence such as "*The ground is frozen, and besides the snow has covered everything.*"

In the crowdsourcing task we designed, the annotators were required to answer 2 questions for a given sentence. In the first question, they were asked to detect the human senses conveyed or directly described by a given sentence. To exemplify these cases, we provided two examples such as "*I saw the cat*" that directly mentions the action of seeing and "*The sun was shining on the blue water.*" that conveys the sense of sight by using visual descriptions or elements like "*blue*" or "*shine*" which are notable for their visual properties. The annotators were able to select more than one sense for each sentence and together with the five senses we provided another option as *None* which should be selected when an annotator could not associate a sentence with any sense. The second question was devoted do determining word-sense associations. Here, the annotators were expected to associate the words in each sentence with at least one sense. Again, annotators could choose *None* for every word that they could not confidently associate with a sense.

The reliability of the annotators was evaluated on the basis of 20 control sentences which were highly associated with a specific sense and which included at least one sensorial word. For instance, for the control sentence "*The skin you love to touch*", we only considered as reliable the annotators who associated the sentence with *touch* and the word *touch* with the sense *touch*[6]. Similarly, for the slogan "*The most colourful name in cosmetics.*", an annotator was expected to associate the sentence with at least the sense *sight* and the word *colorful* to at least the sense *sight*. The raters who scored at least 70% accuracy on average on

[6]If the annotators gave additional answers to the expected ones, we considered their answers as correct.

the control questions for the two tasks were considered to be reliable. Each unit was annotated by at least 10 reliable raters.

Similarly to Mohammad (2011) and Özbal et al. (2011), we calculated the majority class of each annotated item to measure the agreement among the annotators. Table 2 demonstrates the observed agreement at both word and sentence level. Since 10 annotators participated in the task, the annotations with a majority class greater than 5 can be considered as reliable (Özbal et al., 2011). Indeed, for 85.10% of the word annotations the absolute majority agreed on the same decision, while 77.58% of the annotations in the sentence level have majority class greater than 5. The high agreement observed among the annotators in both cases confirms the quality of the resulting gold standard data.

In Table 3, we present the results of the annotation task by providing the association percentage of each category with each sense, namely *sight (Si)*, *hear (He)*, *taste (Ta)*, *smell (Sm)* and *touch (To)*. As demonstrated in the table, while the sense of *sight* can be observed in almost every advertisement category and in *story*, *smell* and *taste* are very rare. We observe that the story sentences invoke all sensory modalities except *taste*, although the percentage of sentences annotated with *smell* is relatively low. Similarly, *personal care* category has an association with four of the senses while the other categories have either very low or no association with some of the sense classes. Indeed, the perceived sensorial effects in the sentences vary according to the category such that the slogans in the *travel* category are highly associated with *sight* whereas the *communication* category is highly associated with *hearing*. While the connection of the *food* and *beverages* categories with *taste* is very high as expected, they have no association with the sense of *smell*. This kind of analysis could be useful for copywriters to decide which sensory modalities to invoke while creating a slogan for a specific product category.

Category	Si	He	Ta	Sm	To
personal care	49.36	10.75	0.00	13.29	26.58
travel	58.18	0.00	29.09	0.00	12.72
fashion	43.47	0.00	0.00	26.08	30.43
beauty	84.56	0.00	0.00	0.00	15.43
computing	32.25	59.13	0.00	0.00	8.60
food	0.00	5.46	94.53	0.00	0.00
beverages	22.68	0.00	59.79	0.00	17.52
communications	25.00	67.50	0.00	0.00	0.075
electronics	45.94	54.05	0.00	0.00	0.00
education	28.57	42.85	0.00	0.00	28.57
transport	61.81	38.18	0.00	0.00	0.00
story	58.37	20.81	0.00	7.23	13.57

Table 3: The categories of the annotated data and their sense association percentages.

4.2 Evaluation Measures

Based on the annotation results of our crowdsourcing task, we propose an evaluation technique considering that a lemma-POS or a sentence might be associated with more than one sensory modalities. Similar to the evaluation framework defined by Özbal et al. (2011), we adapt the evaluation measures of SemEval-2007 English Lexical Substitution Task (McCarthy and Navigli, 2007), where a system generates one or more possible substitutions for a target word in a sentence preserving its meaning.

For a given lemma-POS or a sentence, which we will name as *item* in the rest of the section, we allow our system to provide as many sensorial associations as it determines by using a specific lexicon. While evaluating a sense-item association of a method, a *best* and an *oot* score are calculated by considering the number of the annotators who associate that sense with the given item, the number of the annotators who associate any sense with the given item and the number of the senses the system gives as an answer for that item. More specifically, *best* scoring provides a credit for the best answer for a given item by dividing it to the number of the answers of the system. *oot* scoring, on the other hand, considers only a certain number of system answers for a given item and does not divide the credit to the total number of the answers. Unlike the lexical substitution task, a limited set of labels (i.e., 5 sense labels and *none*) are allowed for the sensorial annotation of sentences or lemma-POS pairs. For this reason, we reformulate *out-of-ten (oot)* scoring used by McCarthy and Navigli (2007) as out-of-two.

In Equation 3, *best* score for a given item *i* from the set of items *I*, which consists of the items annotated with a specific sense by a majority of 5 annotators, is formulated where H_i is the multiset of gold standard sense associations for item *i* and S_i is the set of sense associations provided by the system. *oot* scoring, as formulated in Equation 4, accepts up to 2 sense associations s from the answers of system S_i for a given item *i* and the credit is not divided by the number of the answers of the

system.

$$best\,(i) = \frac{\sum_{s \in S_i} freq\,(s \in H_i)}{|H_i| \cdot |S_i|} \quad (3)$$

$$oot\,(i) = \frac{\sum_{s \in S_i} freq\,(s \in H_i)}{|H_i|} \quad (4)$$

As formulated in Equation 5, to calculate the precision of an item-sense association task with a specific method, the sum of the scores (i.e., *best* or *oot*) for each item is divided by the number of items *A*, for which the method can provide an answer. In recall, the denominator is the number of the items in the gold standard for which an answer is given by the annotators.

$$P = \frac{\sum_{i \in A} score_i}{|A|} \qquad R = \frac{\sum_{i \in I} score_i}{|I|} \quad (5)$$

4.3 Evaluation Method

For the evaluation, we compare the accuracy of a simple classifier based on Sensicon against two baselines on a sense classification task both at word and sentence level. To achieve that, we use the gold standard that we obtain from the crowdsourcing task and the evaluation measures *best* and *oot*. The lexicon-based classifier simply assigns to each word in a sentence the sense values found in the lexicon. The first baseline assigns the most frequently annotated sensory modality, which is *sight*, via crowdsourcing task with a float value of 1.0 to each lemma-POS pair in the sensorial lexicon. The second baseline instead builds the associations by using a Latent Semantic Analysis space generated from the same subset of LDC that we exploit for constructing Sensicon. More specifically, this baseline calculates the LSA similarities between each candidate lemma-POS pair and sense class by taking the cosine similarity between the vector of the target lemma-POS pair and the average of the vectors of the related sensory word (i.e., *see*, *hear*, *touch*, *taste*, and *smell*) for each possible POS tag. For instance, to get the association score of a lemma-POS pair with the sense sight, we first average the vectors of see (noun) and see (verb) before calculating its cosine similarity with the target lemma-POS pair.

For the first experiment, i.e., word-sense association, we automatically associate the lemma-POS pairs obtained from the annotated dataset with senses by using i) Sensicon, ii) the most-frequent-sense baseline (MFS), iii) the LSA baseline. To achieve that, we lemmatize and POS tag each sentence in the dataset by using Stanford Core NLP. In the end, for each method and target word, we obtain a list of senses sorted according to their sensorial association values in decreasing order. It is worth noting that we only consider the non-negative sensorial associations for Sensicon and both baselines. For instance, Sensicon associates the noun *wine* with [*smell, taste, sight*]. In this experiment, *best* scoring considers the associated senses as the best answer, *smell, taste, sight* according to the previous example, and calculates a score with respect to the best answer in the gold standard and the number of the senses in this answer. Instead, *oot* scoring takes the first two answers, *smell* and *taste* according to the previous example, and assigns the score accordingly.

To determine the senses associated with a sentence for the second experiment, we use a method similar to the one proposed by Turney (2002). For each sense, we simply calculate the average score of the lemma-POS pairs in a sentence. We set a threshold value of 0 to decide whether a sentence is associated with a given sense. In this manner, we obtain a sorted list of average sensory scores for each sentence according to the three methods. For instance, the classifier based on Sensicon associates the sentence *Smash it to pieces, love it to bits.* with [*touch*, *taste*]. For the *best* score, only *touch* would be considered, whereas *oot* would consider both *touch* and *taste*.

4.4 Evaluation Results

In Table 4, we list the F1 values that we obtained with the classifier using Sensicon and the two baselines (MFS and LSA) according to both *best* and *oot* measures. In addition, we provide the performance of Sensicon in two preliminary steps, before bootstrapping (BB) and after bootstrapping (AB) to observe the incremental progress of the lexicon construction method. As can be observed from the table, the best performance for both experiments is achieved by Sensicon when compared against the baselines.

While in the first experiment the lexicon generated after the bootstrapping step (AB) provides a very similar performance to the final lexicon according to the *best* measure, it can only build sense associations for 69 lemmas out of 153 appearing in the gold standard. Instead, the final lexicon attempts to resolve 129 lemma-sense associations and results in a better recall value. Additionally, AB yields a very high precision as expected,

since it is created by a controlled semantical expansion from manually annotated sensorial words. BB lexicon includes only 573 lemmas which are collected from 277 synsets and we can not obtain 2 sense association scores for *oot* in this lexicon since each lemma is associated with only one sense with a value of 1. The LSA baseline yields a very low performance in the *best* measure due to its tendency to derive positive values for all sensorial associations of a given lemma-POS tuple. Another observed shortcoming of LSA is its failure to correlate the names of the colors with *sight* while this association is explicit for the annotators. On the other hand, LSA baseline significantly improves the MFS baseline with a p-value of 0.0009 in *oot* measures. This result points out that even though LSA provides very similar positive association values for almost all the sensory modalities for a given item, the first two sensorial associations with the highest values yield a better performance on guessing the sensorial characteristics of a lemma-POS. Nevertheless, Sensicon significantly outperforms the LSA baseline in both *best* and *oot* measures with the p-values of 0.0009 and 0.0189 respectively. The statistical significance tests are conducted using one-sided bootstrap resampling (Efron and Tibshirani, 1994).

Concerning the sentence classification experiment, the classifier using Sensicon yields the highest performance in both measures. The very high F1 value obtained with the *oot* scoring indicates that the right answer for a sentence is included in the first two decisions in many cases. Sensicon significantly outperforms the LSA baseline on the *best* measure (p-value = 0.0069). On the other hand, when systems are allowed to provide two answers (*oot*), the performance of LSA comes close to Sensicon in terms of F1 measure.

After the manual analysis of Sensicon and gold standard data, we observe that the sensorial classification task could be nontrivial. For instance, a story sentence "*He went to sleep again and snored until the windows shook.*" has been most frequently annotated as *hearing*. While the sensorial-lexicon classifier associates this sentence with *touch* as the best answer, it can provide the correct association *hearing* as the second best answer. To find out the best sensorial association for a sentence, a classification method which exploits various aspects of sensorial elements in a sentence, such as the number of sensorial words or their dependencies, could be a better approach than using only the average sensorial values.

	Lemma		Sentence	
Model	*best*	*oot*	*best*	*oot*
Most-Frequent-Sense	33.33	33.33	38.90	38.90
LSA	18.80	70.38	53.44	76.51
Lexicon-BB	45.22	45.22	49.60	51.12
Lexicon-AB	55.85	55.85	59.89	63.21
Sensicon	55.86	80.13	69.76	80.73

Table 4: Evaluation results.

Based on our observations of the error cases, we believe that synaesthesia, which is one of the most common metaphoric transfers in language (Williams, 1976), should be further explored for sense classification. As an example observation, the advertisement slogan "*100% pure squeezed sunshine*" is associated with *touch* as the best answer by Sensicon and *taste* by LSA baseline while it is most frequently annotated as *sight* in the gold standard. This slogan is an example usage of synaesthesia and metaphors in advertising language. To clarify, a product from the category of *beverages*, which might be assumed to have a *taste* association, is described by a metaphorical substitution of a *taste*-related noun, most probably the name of a fruit, with a *sight*-related noun; *sunshine*. This metaphorical substitution, then used as the object of a *touch*-related verb, *to squeeze*, produces a synaesthetic expression with *touch* and *sight*.

5 Conclusion

In this paper we have presented the construction of Sensicon, a sensorial lexicon, which associates words with sensory modalities. This novel aspect of word semantics is captured by employing a two-step strategy. First, we collected seed words by using a bootstrapping approach based on a set of WordNet relations. Then, we performed a corpus based statistical analysis to produce the final lexicon. Sensicon consists of 22,684 lemma-POS pairs and their association degrees with five sensory modalities. To the best of our knowledge, this is the first systematic attempt to build a sensorial lexicon and we believe that our contribution constitutes a valid starting point for the community to consider sensorial information conveyed by text as a feature for various tasks and applications. The results that we obtain by comparing our lexicon against the gold standard and two baselines are promising even though not conclusive. The results confirm the soundness of the proposed approach for the construction of the lexicon and the useful-

ness of the resource for text classification and possibly other computational applications.

Sensicon is publicly available upon request to the authors so that the community can benefit from it for relevant tasks. From a resource point of view, we would like to explore the effect of using different kinds of WordNet relations during the bootstrapping phase. It would also be interesting to experiment with relations provided by other resources such as ConceptNet (Liu and Singh, 2004), which is a semantic network containing common sense, cultural and scientific knowledge. We would also like to use the sensorial lexicon for various applicative scenarios such as slanting existing text towards a specific sense with text modification. We believe that our resource could be extremely useful for automatic content personalization according to user profiles. As an example, one can imagine a system that automatically replaces hearing based expressions with sight based ones in pieces of texts for a hearing-impaired person. Automating the task of building sensorial associations could also be beneficial for various tasks that need linguistic creativity. For instance, copywriters can take advantage of a system detecting the sensorial load of a piece of text to generate more appropriate advertisement slogans for specific product categories. Finally, we plan to investigate the impact of using sensory information for metaphor detection and interpretation based on our observations during the evaluation. For instance, the synaesthetic metaphor *bittersweet symphony* could be detected by determining the sensorial characterizations of its components.

Acknowledgements

We would like to thank Daniele Pighin for his insightful comments and valuable suggestions.
This work was partially supported by the PerTe project (Trento RISE).

References

Ebba Cecilia Ovesdotter Alm. 2008. *Affect in Text and Speech*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. pages 86–90. Association for Computational Linguistics.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *LREC*.

Douglas A. Bernstein. 2010. *Essentials of Psychology*. PSY 113 General Psychology Series. Cengage Learning.

Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. 1992. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual Workshop on Computational learning theory*.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505.

Ivan E. de Araujo, Edmund T. Rolls, Maria Inés Velazco, Christian Margot, and Isabelle Cayeux. 2005. Cognitive modulation of olfactory processing. *Neuron*, 46(4):671–679.

Gaël Harry Dias, Mohammed Hasanuzzaman, Stéphane Ferrari, and Yann Mathet. 2014. Tempowordnet for sentence time tagging. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 833–838, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Bradley Efron and Robert J. Tibshirani. 1994. *An introduction to the bootstrap*, volume 57. CRC press.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Lushan Han, Tim Finin, Paul McNamee, Anupam Joshi, and Yelena Yesha. 2013. Improving word similarity by augmenting pmi with estimates of word polysemy. *Knowledge and Data Engineering, IEEE Transactions on*, 25(6):1307–1322.

Thomas Kjeller Johansen. 1997. *Aristotle on the Sense-organs*. Cambridge University Press.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Hugo Liu and Push Singh. 2004. Conceptnet - a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, October.

Asifa Majid and Stephen C. Levinson. 2011. The senses in language and culture. *The Senses and Society*, 6(1):5–18.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34. Association for Computational Linguistics.

Saif M. Mohammad. 2011. Colourful language: Measuring word-colour associations. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 97–106. Association for Computational Linguistics.

Gözde Özbal, Carlo Strapparava, Rada Mihalcea, and Daniele Pighin. 2011. A comparison of unsupervised methods to associate colors with words. In *Affective Computing and Intelligent Interaction*, pages 42–51. Springer.

Jessica Perrie, Aminul Islam, Evangelos Milios, and Vlado Keselj. 2013. Using google n-grams to expand word-emotion association lexicon. In *Computational Linguistics and Intelligent Text Processing*, pages 137–148. Springer.

Mario Pricken. 2008. *Creative Advertising Ideas and Techniques from the World's Best Campaigns*. Thames & Hudson, 2^{nd} edition.

Raul Rodriguez-Esteban and Andrey Rzhetsky. 2008. Six senses in the literature. The bleak sensory landscape of biomedical texts. *EMBO reports*, 9(3):212–215, March.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proceedings of LREC*, volume 4, pages 1083–1086.

Sara Tonelli and Daniele Pighin. 2009. New features for framenet - wordnet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09)*, Boulder, CO, USA.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, pages 680–690.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Alessandro Valitutti, Carlo Strapparava, and Oliviero Stock. 2004. Developing affective lexical resources. *PsychNology Journal*, 2(1):61–83.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. Wiley-Interscience.

Joseph M. Williams. 1976. Synaesthetic adjectives: A possible law of semantic change. *Language*, pages 461–478.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136. Association for Computational Linguistics.