Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2991–3000
Marseille, 11–16 May 2020

Building Semantic Grams of Human Knowledge

Valentina Leone[1], Giovanni Siragusa[1], Luigi Di Caro[1], Roberto Navigli[2]
[1]Computer Science Department, University of Turin, Italy
[2] Computer Science Department, Sapienza University of Rome, Italy
{leone, siragusa, dicaro}@di.unito.it
navigli@di.uniroma1.it

Abstract

Word senses are typically defined with textual definitions for human consumption and, in computational lexicons, put in context via lexical-semantic relations such as synonymy, antonymy, hypernymy, etc. In this paper we embrace a radically different paradigm that provides a slot-filler structure, called "semagram", to define the meaning of words in terms of their prototypical semantic information. We propose a semagram-based knowledge model composed of 26 semantic relationships which integrates features from a range of different sources, such as computational lexicons and property norms. We describe an annotation exercise regarding 50 concepts over 10 different categories and put forward different automated approaches for extending the semagram base to thousands of concepts. We finally evaluate the impact of the proposed resource on a semantic similarity task, showing significant improvements over state-of-the-art word embeddings. We release the complete semagram base and other data at `http://nlp.uniroma1.it/semagrams`.

Keywords: Semagrams, word senses, concept representation, lexical semantics

1. Introduction

The representation of knowledge is one of the great dreams of Artificial Intelligence. Acquiring and encoding knowledge is essential not only to improve historical NLP tasks such as Word Sense Disambiguation and Machine Translation, but also to enable numerous applications (Hovy et al., 2013) like intelligent personal assistants, Question Answering, Information Retrieval, etc. Lexical-semantic knowledge has typically been encoded on a large scale using word senses as meaning units, starting from WordNet (Miller, 1995) and then carrying on with VerbNet (Schuler, 2005), PropBank (Kingsbury and Palmer, 2002) and, more recently, VerbAtlas (Di Fabio et al., 2019). Large repositories of frames have been introduced with FrameNet (Baker et al., 1998) and its counterparts in other languages. However, frames are focused on situation-based representations with semantic roles and commonly involved objects, without defining the embodied ontological semantics of single concepts.

Word senses, on the other hand, suffer from a lack of explicit common-sense semantic information. Indeed, the well-known fine granularity of word senses in WordNet (Palmer et al., 2007) is due to the lack of a meaning encoding system capable of managing the representation of concepts in a flexible way. To address this issue, Pustejovsky introduced an innovative model (Pustejovsky, 1991) based on qualia roles to enable the creation of new meanings via semantic slot filling. However, the approach was limited by the number and type of these slots.

An interesting and novel extension was presented by Moerdijk et al. (2008) with the ANW dictionary and the introduction of the concept of *semagram*. A semagram is a conceptual structure that describes a lexical entity on the basis of a wide range of characteristics, defined with a rich slot-filler structure. The semagrams provided in the ANW dictionary are, however, limited in coverage, often expressed with a fragmented set of semantic slots and written in Dutch.

The aim of this paper is threefold: *(i)* to propose a novel model of semantic representation of concepts; *(ii)* to create a new semantic resource through manual annotation and semi-automatic techniques; and *(iii)* to evaluate the impact of the resource on a text similarity task, in comparison with state-of-the-art word and sense embeddings. More in detail, we provide the following contributions:

1. a new approach to semagrams by bringing together advancements going in the same direction as Property Norms (McRae et al., 2005; Devereux et al., 2014);
2. a manual annotation exercise on 50 concepts belonging to 10 different categories; as a result, we have built a total of 1,621 manually disambiguated slot-filler instances with a set of annotation guidelines;
3. a semi-automated extension strategy involving Sketch Engine (Kilgarriff et al., 2014) and word2vec embeddings (Mikolov et al., 2013b) that led to the annotation of 250 additional concepts in 1/30 of the original required time (i.e., 115 seconds per concept on average, instead of 57 minutes);
4. an automatic extension based on a novel notion of *semantic profile* and the automatic learning of abstract lexical-syntactic patterns from Wikipedia, that generated thousands of semagram annotations with good precision values;
5. the notion of *semantic propagation* as the way, enabled by the model, to propagate individual semantic properties through a taxonomy of word senses; this allowed the automatic extension of the resource from the initial 50 concepts to 923 additional hyponym concepts with an accuracy of 85.43% calculated through a manual validation on a sample of 400 slot-filler pairs.
6. a test of the model significance in a semantic similarity task, outperforming all state-of-the-art embedding models while giving lexicalized meaning to similarity values.

2. Related Work

Models of explicit lexical-semantic knowledge representation may be classified into five main categories, which we overview in this Section.

Computational Lexicons Word senses are the basis of computational lexicons such as WordNet (Miller, 1995) and its counterparts in other languages (Bond and Foster, 2013). They usually provide human-readable concept definitions and contextualize meanings mainly in terms of the paradigmatic relations (hypernymy, meronymy) that hold between them. While larger resources such as BabelNet (Navigli and Ponzetto, 2010) also integrate other relations, these are typically unlabeled and are not systematized. A key unsolved issue with wordnets is the fine granularity of their inventories.

Frames Frames (Fillmore, 1977) encode meanings through simple slot-filler structures. In other words, they represent knowledge as a set of attributes and values with the aim of defining situations or events. However, although semantic roles provide prototypical lexical units, these latter are not the primary focus of the frame model. Therefore, frames are not a viable option for encoding the prototypical meaning of concepts.

Corpus-based Models Corpus-based semantic models are based on the idea that similar words are used in similar contexts (Harris, 1954). Corpus Pattern Analysis (Hanks, 2004, CPA) is a procedure for lexicographers to map meaning to words based on the theory of Norms and Exploitations (Hanks, 2013). The underlying idea is to analyse the prototypical syntagmatic patterns of words in their use in large corpora. CPA focuses on patterns for nouns and for verbs and their argument structure, and is centered on the lexical aspects of meaning. CPA inherits some of its intents from the Generative Lexicon (Pustejovsky, 1991), the theory of preference semantics (Wilks, 1975), and others. Generally speaking, words are not taken in isolation and the meaning they are attributed is ascertained on a contextual basis through prototypical sentence patterns. However, these theories and methods for building semantic resources remain linked to the lexical basis and are suitable for the manual effort of lexicographers. On the other hand, automatic approaches such as those of Almuhareb and Poesio (2004), Baroni et al. (2010), Navigli and Velardi (2010) and Boella and Di Caro (2013) used surface text patterns to automatically extract concept descriptions. However, these methods do not have a knowledge model as they extract semantics from statistically significant word properties. Mishra et al. (2017) extract domain-targeted knowledge, identifying clusters of similar-meaning predicates. However, the approach does not organize knowledge through semantic relations, and is not suitable for general knowledge building. Finally, Open Information Extraction (OIE) achieved notable results in extracting relational phrases from large corpora such as Wikipedia and the Web (Banko et al., 2007; Wu and Weld, 2010; Carlson et al., 2010; Fader et al., 2011; Del Corro and Gemulla, 2013).

Common-sense Knowledge Common-sense knowledge (CSK) may be described as a set of shared and general facts or views of a set of concepts. CSK displays some similarity with a semagram-type of knowledge in that it describes the kind of general information that humans use to describe, differentiate and reason about the conceptualizations they have in mind. ConceptNet (Speer and Havasi, 2012; Speer et al., 2016) is one of the largest resources of this kind, collecting and automatically integrating data starting from the original MIT Open Mind Common Sense project[1]. However, terms in ConceptNet are not disambiguated, which leads to the confusion of lexical-semantic relations involving concepts denoted by ambiguous words (e.g. *mouse* as a device vs. a rodent). NELL (Carlson et al., 2010) matches entity pairs from seeds to extract relational phrases from a Web corpus, although without linking patterns to a slot-filler knowledge model, being mostly oriented to named entities rather than concept descriptions. Property norms (McRae et al., 2005; Devereux et al., 2014) represent a similar kind of resource, which is more focused on the cognitive and perception-based aspects of word meaning. Norms, in contrast with ConceptNet, are based on empirically constructed semantic features via questionnaires asking people to produce features they consider important for some target concept (e.g., a *crocodile* is often associated with the norm *is-dangerous*). The problem with norms is that they mix properties and values (i.e., slots and fillers), and they do not represent complete descriptions (usually, only immediate and common-sense facts are reported).

Geometric Approaches to Semantics A radically different approach is based on vector space models of lexical (sometimes semantic) representations. Conceptual Spaces (Gärdenfors, 2004) provide a geometric approach to meaning, viewing concepts as vectors whose dimensions are qualitative features. For example, colors may be represented with three dimensions: hue, saturation, and brightness. While this model allows for direct similarity computation among instances, the knowledge it encodes does not define concepts explicitly, and dimensions usually represent perceptual mechanisms only. Other methods include topic models such as Latent Semantic Analysis (Dumais, 2004), Latent Dirichlet Allocation (Blei et al., 2003), and, more recently, embeddings of words (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2016) and word senses (Huang et al., 2012; Iacobacci et al., 2015; Scarlini et al., 2020). However, the relations holding between vector representations are not typed, nor are they organized systematically.

Semagrams To address the issues with existing approaches to concept representation, Moerdijk et al. (2008) proposed the concept of semantic gram, or *semagram*, and manually constructed a semantic resource in Dutch – the ANW (Algemeen Nederlands Woorden boek - General Dutch Dictionary) Dictionary – containing approximately 70,000 headwords, organized in 20 domains (animal, plants, etc.) with a total of 200 slots. However, most of the headwords lack any semagram annotation and only a few hundred provide rich annotations. Moreover, the guidelines lack a formal description: lexicographers were only invited to characterise features in terms of short statements

[1] `http://media.mit.edu/research/groups/5994/open-mind-common-sense`

about the headword and could autonomously decide what was relevant and appropriate for inclusion.

The main goal of this paper is to revise the knowledge model and systematize and automatize the creation and annotation of semagrams.

3. The Semagram Knowledge Model

A semagram is a flexible structure for encoding the semantics of a given concept via a slot-filler structure. In Table 1 we show an excerpt of semagram for the concepts of *dog* (right) and *piano* (left). As can be seen, the two semagrams share several slot types (but, obviously, not their values), while some slots are used only for one of the two concepts due to their belonging to different categories.

The first contribution of this paper is a careful, systematic analysis, unification and extension of semagram slots from different resources. This analysis was initially performed on a development sample of 20 concepts over the 10 different categories used by Silberer et al. (2013) and then refined using another set of 30 concepts (3 per category). We chose these categories for their variety across conceptual types. The full sample is shown in Table 2 (development concepts are shown in bold in each category row). The 5 concepts in each category were chosen from among popular terms according to a diversification criterion specific to each category that would maximize the intra-category variability. For example, for the category *Animals*, we chose five concepts with different kinds of movement, for the category *Clothes* we selected those that are worn on different body parts, etc. Each concept and each filler was associated with a single WordNet synset.

The objective of our work was to define a wide-coverage semagram knowledge model in terms of the semagram slots that would be used to cover all the facets of a concept description. To do this, we followed the basic steps below:

- we started by analyzing concepts belonging to one category at a time (from *C1* to *C10* as in Table 2);
- we translated each concept into Dutch and collected all its semagrams available in the ANW dictionary (Moerdijk et al., 2008);
- we created clusters of ANW slots that shared identical or very similar fillers (e.g., the *habitat*, *place* and *location* slots used to contain identical or often similar fillers across concepts of the same category);
- for the same concepts, we then collected the semantic features contained in the Property Norms (Devereux et al., 2014), adding in some cases new semagram slots that were not covered by the ANW dictionary (e.g., *consistency* for concepts belonging to the category *food*);
- we repeated the previous step with the Visual Attributes (Silberer et al., 2013);
- we finally added new semagram slots, encoding relevant prototypical semantic information that was not found in the above-mentioned resources.

As a result of applying the above steps to the development sample of 20 concepts we obtained an initial set of 24 slots. We then refined our semagram knowledge model by annotating the remaining 30 concepts (3 per category) in our sample. In this last step we were able to annotate the new concepts without revising the knowledge model, except for adding two new slots: *bodyPart* (which helped in encoding parts of the body involved in the use of objects, especially in vehicles, home objects, instruments, artifacts, and tools) and *howToUse* (actions for using objects such as instruments, artifacts and tools). The resulting semagram base contains a total of 1,621 manual slot-filler annotations, and 906 distinct fillers across all semagram slots. The complete list of semagram slots with the corresponding textual descriptions is shown in Table 3. To get an idea of the importance of our unification work, Table 4 shows the sources from which each of the 26 slots was derived.

To summarize, the resulting semagram knowledge model differs from the one of its original proposers in several aspects: 1) we defined an XML annotation scheme and the annotation guidelines, uploaded with the resource; 2) we removed, added and merged semagram slots starting from (Moerdijk et al., 2008) and integrating features from Property Norms (Devereux et al., 2014) and Visual Attributes (Silberer et al., 2013), as detailed above; 3) fillers were disambiguated and put in their lemma form in a comma-separated list (cf. the two right columns in Table 1).

4. Extending the Semagram Base

In this section, we describe three strategies for the extension of the initial semagram base:

1. a semi-automated approach based on Sketch Engine (Kilgarriff et al., 2014) and word2vec embeddings (Mikolov et al., 2013b);
2. an automatic technique based on the learning of syntactic patterns coupled with an abstraction step relying on the notion of *semantic profiles*;
3. a second automatism, called *semantic propagation*, which exploits the WordNet hypernym relations to propagate single slot-filler pairs.

4.1. A Semi-Automated Approach

By having 5 concepts for each of the 10 categories, we initially looked for a further 25 new concepts per category, with the objective of extending the semagram base to 300 concepts from the initial set of 50. This search for additional concepts was made manually, by browsing the WordNet structure (e.g., sister terms, hyponyms, etc.).

First, for each concept c in the original set, we created links between its semagram slots and the grammar relations of c in Sketch Engine (SE, from now on) (Kilgarriff et al., 2014) based on the maximum number of matching words/fillers. For example, the slot "*activity*" of concept "*dog*" was linked through 4 fillers (e.g., "*bark*", "*walk*", "*chase*", "*bite*") with the SketchEngine grammar relation "*verbs with X as subject*".

After the linking stage, all words within the linked SE grammar relations were thus automatically added as fillers

Slot	**Concept** *Piano*	**Concept** *Dog*
GENERALIZATION	musical instrument#N	mammal#N; animal#N
SPECIALIZATION	grand piano#N; upright piano#N	Boxer#N; Reottweiler#N; Poodle#N; Yorkshire-Terrier#N; Bulldog#N; Beagle#N; Golden-Retriever#N; German-Shepherds#N; Labrador#N
COLORPATTERN	black#A|N; white#A|N; brown#A|N	solid#A; spotted#A; grey#A|N; black#A|N; brown#A|N; white#A|N
PART	eighty-eight black,white key#N; lid#N lid-prop#N; music-rack#N; fall-board#N; three leg#N; three pedal#N; metal string#N; hammer#N; soundboard#N; case#N	coat#N; fur#N; hair#N; forty-two tooth#N; four sharp fang#N; long,short tail#N; four leg#N; two eye#N; bone#N; tongue#N; two ear#N; flash#N; claw#N; mouth#N; head#N; whisker#N; jaw#N; neck#N; nose#N; four paw#N; pad#N
MATERIAL	wood#N; ivory#N; metal#N; felt#N	
PURPOSE	play#V; accompaniment#N; solo#N	pet#N; guard#N|V; hunting#N; guide-animal#N; research#N
USER	pianist#N; musician#N; orchestra#N	shepherd#N; hunter#N; blind#N; policeman#N; trainer#N
HOWTOUSE	sit#V; press#N	
BODYPART	finger#N; foot#N	
ACCESSORY	seat#N	collar#N; muzzle#N; leash#N; kennel#N
PLACE		kennel#N
MOVEMENT		run#V; walk#V
SOUND		bark#N; yelp#N; growl#N; whining#N
ACTIVITY		walk#V; run#V; eat#V; drink#V; bite#V; chew#V; bury#V; fetch#V, play#V; breathe#V; bark#V; yelp#V; growl#V

Table 1: An excerpt of two semagrams for the concepts of *dog* and *piano*.

Category	**Selected concepts**
C1: *animals*	**bee**, **dog**, elephant, snail, frog
C2: *food*	**apple**, **carrot**, corn, bread, wine
C3: *vehicles*	**airplane**, **car**, bicycle, ship, tractor
C4: *clothes*	**skirt**, **boot**, glove, cap, scarf
C5: *home*	**mug**, **spoon**, sink, rocker, gate
C6: *appliance*	**projector**, **telephone**, thermometer, dishwasher, stove
C7: *instruments*	**accordion**, **cello**, clarinet, drum, piano
C8: *artifacts*	**helmet**, **bracelet**, mirror, umbrella, typewriter
C9: *tools*	**anchor**, **hoe**, scissors, screws, tongs
C10: *containers*	**bag**, **barrel**, bottle, bucket, tray

Table 2: The sample of concepts selected for manual annotation (in **bold** the concepts from the development sample), with their corresponding category.

in the original semagrams[2] (if not already present). In the above example, the word "*lick*" in SE has been automatically disambiguated and added to the semagram of "*dog*'.

The method allowed the initial manual annotation to be enriched from 1,621 (slot, filler) pairs to 1,913 (292 new pairs had not been found by the annotators) for the initial 50 concepts, after a manual check for correctness.

Then, the same procedure was applied to the extension set of 250 new concepts, by considering the words (i.e., candidate fillers) in the previously-linked SE grammar relations. The final result was a total of 6,701 (slot, filler) pairs. The manual check of the automatically retrieved (slot, filler) annotations required around 1/30 of the average "from-scratch" annotation time (i.e., 115 seconds per concept on average, instead of 57 minutes). This was due to multiple factors: 1) the selection of correct (slot, filler) pairs is simpler than their search from different sources; 2) words are automatically POS-tagged; 3) disambiguated[3]; and 4) already organized in semagram slots. This approach is, however, valid on an extension set built around manually-annotated seed concepts of a particular category.

4.2. An Automatic Approach based on Semantic Profiles

Using the WordNet synset identifier associated with each filler of the semagram base, we built the distributional semantic profile for each slot retrieving the most frequent WordNet *supersense*[4] for each of them. We built different semantic profiles for each of the ten categories in order to make them more precise. Since adjectives are not semantically well categorised in WordNet (they often have

[2] A filtering step based on word2vec embeddings (Mikolov et al., 2013b) has been applied to remove those fillers having a similarity lower than 0.2 with the embedding of the lexicalized concept.

[3] We utilized Babelfy (Moro et al., 2014) as Word Sense Disambiguation tool - `http://babelfy.org`.

[4] `https://wordnet.princeton.edu/man/lexnames.5WN.html`.

Slot	Description
S1: accessory	All those objects that may have to do with X. The constraint is that there must be a physical contact and that the use of such object is strictly necessary for X.
S2: activity	All actions that X can actively or consciously do.
S3: behavior	All the psychological features of X, including they attitude to they nature.
S4: bodyPart	All the body parts which are involved in interacting with X.
S5: colorPattern	All the features that refer to the color or texture of X.
S6: consistency	All the entries with which the noticeable to the touch consistency or texture of X can be described.
S7: content	All the entities which might be contained within X, without being constitutive parts of it.
S8: efficiency	Positive (efficiency) or negative (inefficiency) features of X related to their function.
S9: generalization	Classification of X related to hypernyms.
S10: group	Names that indicates a group of animals of the same species of X.
S11: howToUse	All the actions or states required to operate, employ, interact with or perceive the existence of X.
S12: material	Material of which X is composed.
S13: movement	Terms that describe the type and speed of movement.
S14: part	All the constitutive parts of X.
S15: place	All the entities in which X can be experienced, found or perceived.
S16: product	All types of entity that can be derived from X through its processing or through natural processes.
S17: purpose	All of the purposes for which X is interacted with.
S18: shape	Form of X.
S19: size	Size of X.
S20: smell	All the entries with which the smell of X can be described.
S21: sound	All the entries with which the sound of X can be described.
S22: specialization	Classification of X in terms of their hyponyms.
S23: supply	The power mode that allows the functioning of X.
S24: taste	Contains information on the taste of a food.
S25: time	All the entries which link X with the time flow or with specific moments of time.
S26: user	All the kinds of living beings which are able to operate, employ, interact with or perceive X.

Table 3: The semagram knowledge model in terms of the slots resulting from our sample concept annotation. By X we refer to a general concept, chosen from among the fifty annotated concepts.

R	**S1**	S2	S3	**S4**	S5	S6	**S7**	**S8**	S9	S10	S11	S12	S13	S14
ANW		x	x		x				x	x	x	x	x	x
PN					x	x			x					x
VA			x		x							x		x

R	S15	S16	S17	S18	S19	S20	S21	S22	**S23**	S24	S25	S26	**-**	-
ANW	x	x	x	x	x	x	x	x		x	x	x		
PN						x	x			x				
VA				x	x									

Table 4: Overview of the knowledge model derivation and integration. Columns are semagram slots in the knowledge model, while rows represent the three sources: the ANW dictionary, the Property Norms (PN), and the Visual Attributes (VA). Cells have been filled when the corresponding resource was used for the given slot, while empty columns indicate new slot proposals from the authors (marked in **bold**).

adj.all as supersense), we kept the nominal synset linked to an adjective via the WordNet relation *semantically-related form* and we chose for the slots whose fillers were mainly adjectives a set of supersenses for the semantically related forms which characterize well their distributional profile (their fillers). Table 5 shows an excerpt of semantic profile of the values for the semagram slot *part* for two different categories (animals and vehicles).

For each of the 50 concepts in our semagram base, we built a corpus by extracting sentences from a semantically-enriched version of Wikipedia (Raganato et al., 2016) which contains the concept together with one of its disambiguated semagram fillers[5]. For example, given the concept

[5] We also considered plural forms, third person, past simple

Animals supersenses	freq.	Values for slot *part*
noun.body	19	'eye', 'tooth', 'flesh', 'belly', 'nail', 'toe', 'mouth', 'abdomen', 'foot', 'tongue', 'bone', 'head', 'nose', 'skin', 'neck', 'jaw', 'ear', 'hair', 'leg'
noun.animal	6	'tail', 'wing', 'paw', 'beak', 'claw', 'feeler'
***Vehicles* supersenses**		
noun.artifact	44	'windshield', 'door', 'saddle', 'gear', 'basket', 'headlights', 'horn', 'window', 'gearbox', 'lifeboat', 'frame', 'taillight', 'roof', 'bell', 'bunk', 'tire', 'brake' ...
noun.communication	2	'pedal', 'radio'

Table 5: Example of semantic distributional profile for slot *part* with categories *Animals* and *Vehicles*.

dog and the slot-value pair (*part*, *tail*), we retrieved those sentences having both the concepts *dog* and *tail* within a window w of words and associated them with the slot *part*. In order to increase the precision of the extraction, we used a limited window $w = 5$. The corpus extracted for our 50-concept semagram base contains 1,040,312 sentences.

The next step consisted of extracting the textual patterns, i.e., for each sentence, the text contained between the concept term and the disambiguated semagram value, on a slot by slot basis. We developed an OIE system, based on (Delli Bovi et al., 2015), to extract phrase excerpts that unveil the relation between a concept and a semagram value. The proposed OIE system takes as input a sentence and two lemmatized arguments (the concept and the semagram value) and returns a phrase excerpt. First, it generates all possible lexical variants of the lemmatized words (the word itself, its plural form, its past simple form, its -ing form, and so forth) by using the Unimorph English Corpus[6] and a set of handcrafted rules. Then, it processes the sentences through two steps: a parsing step and a merging step. In the first step, a Dependency Parser[7] is applied to the sentences to generate a dependency graph. The graph is passed as input to the second phase, where it is merged with the output of a Word Sense Disambiguation (WSD) method which assigns a sense to each word of the sentence according to the context. The OIE system can also modify some edges of the graph to deal with conjunctions (e.g., "and") and coordination (list of words separated by commas). If a node m is connected to a node n with an edge labelled as conjunction (or coordination), the OIE system splits the connection and links m to a neighbour of n which has the edge labelled as object (or modifier).

The output of the system is a set of triples of the form *(argument1, shortest path, argument2)* in which *argument1* and *argument2* are two (possibly disambiguated) words belonging to the lexical variant sets.

Once the patterns are extracted, we search for those that are relevant for a specific semagram slot. Specifically, for each pattern p and slot s, we compute $score(s, p)$ using the following formula:

$$score(s,p) = \frac{freq_s(p)}{\sum_{p' \in Patterns} freq_s(p')} * \frac{1}{H(p)+1} \quad (1)$$

where $freq_s(p)$ is the frequency of the pattern p in the slot s, while $Patterns$ is a set containing all the patterns in s. $H(p)$ is the entropy of the pattern which is based on the distribution of the concepts over the categories. A high value of $score(s, p)$ means that the pattern p is relevant for the slot s.

Thus, finally, after performing each of the aforementioned steps of the OIE system, we manually checked the first 100 top-scored patterns and selected a subset for the automatic extraction of new semagrams. Table 6 shows the selected patterns for semagram slots *material*, *supply*, *colorPattern*, *bodyPart* and *place*.

Slot	Top patterns
colorPattern	was painted, being painted, is painted, were painting, were painted, are painted, are painted in, was painted, ...
supply	runs out of, ran out of, filling with, filled with, fill with, fills with, using, use, uses, used, powered by, ran on, running on, ...
material	make from, making from, makes from, made from, produced from, produce from, producing from, produces from, ...
bodyPart	on, wore on, worn on, wears on, wear on, wearing on around, wears around, wearing around, worn around, ...
place	at, on, in, were built at, was built at, are built at, were at, is at, are at, was at, been at, was lost at, are lost at, ...

Table 6: Top selected patterns for semagram slots *colorPattern*, *supply*, *material*, *bodyPart* and *place*.

We then used two methods for the automatic extension of the semagram base.

Exact match (EM) for sister terms, hyponyms and similar concepts. For each concept c and each slot s in the initial semagram base, we extract the set of sister terms $St(c)$ and hyponyms $Hyp(c)$ using WordNet, and a set $Sim(c)$ of top-10 similar concepts relying on word2vec-GoogleNews-vectors[8] and cosine similarity. Then, we construct queries of the form "$x\ p_s\ y$" where $x \in St(c) \cup Hyp(c) \cup Sim(c)$, p_s is one of the extracted patterns for slot s, and $y \in C_s$, which is the codomain of s (i.e., the existing fillers for that slot in the semagram base);

Wildcards (W) on fillers and concepts. In this case, we construct queries of the form "$c\ p_s\ ?q$", where c is a concept in the semagram base and $?q$ can match any word that follows the pattern in a sentence. We leverage the supersense-

and present continuous for the verbs.

[6] http://www.unimorph.org

[7] We used Mate-Tools parser (http://code.google.com/p/mate-tools).

[8] https://code.google.com/archive/p/word2vec/

based semantic profiles of the semagram values to filter out the retrieved sentences having as filler a supersense which is not contained in the slot profile. Then, we build queries of the form "$?q\ p_s\ y$" where the extracted patterns are concatenated with the fillers and used as queries for extracting new concepts.

Overall, use of the above two methods automatically extracted 4,205 semagrams (from the initial manually-annotated 50, with a multiplication factor of 83x) and a total of 55,245 (slot, filler) pairs (from the 1,621 initial pairs, with a multiplication factor of 34x). The Exact Match (EM) method was able to learn 651 new semagrams, with a total of 1,627 new (slot, filler) pairs. Our experimentation on the challenging wildcard-based strategies (W) allowed the extraction of 2,631 semagrams and 8,786 (slot, filler) pairs.

The last phase concerned the manual validation of the (slot, filler) pairs extracted in the previous Section. We randomly chose 100 pairs for each extraction method, manually evaluating their correctness. Table 7 shows the Precision of the adopted extraction strategies.

Method	Semagrams	Pairs	Avg Precision*
EM $x\ p_s\ y$	651	1,627	85.34%
W $x\ p_s\ ?q$	50	3,742	78.36%
W $?q\ p_s\ y$	2,581	5,044	67.00%

Table 7: Precision of the extraction methods on random samples of 100 slot-filler pairs, starting from the initial manually-annotated semagram base. **Recall cannot be reported as it would mean having the relevant set of slot-filler pairs in the whole Wikipedia for each slot.*

4.3. An Automatic Approach based on Semantic Propagation

Due to the nature of the encoding system, the fine-grained slot-filler pairs represent properties that can be propagated along a taxonomy of word senses. For example, if a car has an engine and a brake, it is likely that a coupé will have them as well. Following this reasoning, we put forward an automatic semantic propagation of single slot-fillers over the WordNet taxonomy. In detail, we automatically propagated all (slot, filler) pairs of the original 50 concepts to all their hyponyms, without relying on any evidence from large corpora. This experiment led to the automatic creation of 923 new semagrams, as hyponyms of the 50 concepts initially annotated, for a total of 44,832 new (slot, filler) pairs. We then manually evaluated the validity of the propagated (slot, filler) pairs on a random set of 400 instances, reaching a precision of 85.43% (these evaluation data will be released with the resource). This test demonstrates two facts: 1) the high presence of semantic redundancy that standard semantic resources do not properly manage with paradigmatic relations and individual glosses; 2) how significant a single (slot, filler) annotation might be in the semagram base, as it can be inherited or propagated through (even more advanced) reasoning processes among different word senses.

5. Evaluation and Impact of Semagrams

The proposed semantic encoding may have impact on several NLP tasks, such as Word Sense Disambiguation and Machine Translation. In this paper, we decided to employ semagrams within the task of word-level semantic similarity, due to their broad range of uses and implications. As mentioned in Section 2., word embeddings often represent a useful source of word-level information as they encode both syntactic and semantic features automatically harvested from large corpora, in accordance with the principles of Distributional Semantics. Their adoption is massive, as demonstrated by their presence and utilization in the most recent scientific literature. Thus, we compare some vectorializations of our semagram base with state-of-the-art embedding models.

5.1. The Evaluation Task

Given a single concept c, the evaluation task first regards the identification of the top-k similar concepts. Then, it checks how many of these fall within the same category $cat(c)$ of concept c. We compared the results of three types of vectorialization of the semagram base with a baseline and four state-of-the-art word embedding models.

5.2. Models at Comparison

In this section, we describe the details of the evaluated models. The concepts considered are the 300 for which we have their semagram representations, organized in the 10 categories of Table 2.

[Concepts-Corpus]. We extracted a corpus from the English Wikipedia[9] made up of the 1,311,124 sentences containing at least one of the 300 concepts in the semagram base. We segmented these sentences by concept, finally applying a tf-idf concept vectorization.

[Semagram-Corpus]. From the above corpus, we selected those sentences (i.e., 1,901) containing at least one filler from the concepts' semagram representations. We then built a feature space of all 3,586 existing (slot, filler) combinations, where the values derive from an adapted tf-idf weighting. In particular, after a process of lemmatization and stopwords removal applied on the corpus, the value of each (slot, filler) dimension is given by the normalized term frequency of the filler in the corpus (i.e., $tf = 1 + log(freq(filler))$) multiplied by an idf score (i.e., how much $filler$ is shared among the concepts).

[Semagram-Binary]. From the semagram base, we constructed a binary model where the feature space is that of the *Semagram-corpus* model. Cells are 1-valued if the concept considered has the related (slot, filler) pairs within its semagram representation. Otherwise, cells are equal to 0.

[Semagram-Mixture]. The features are those of the *above* semagram spaces, while the values come from a weighted-sum of the two (*Semagram-Corpus*, *Semagram-Binary*) matrices. The weight was set to 0.35 and 0.65 respectively.

[GloVe, fastText, ConceptNet] We employed three word embeddings models: the GloVe model presented in (Pennington et al., 2014); the fastText one of (Bojanowski et al., 2016) which integrates subword information in Mikolov's

[9] English Wikipedia dump of November 2014.

model (Mikolov et al., 2013a); and the ConceptNet-based semantic vectors presented in (Speer et al., 2017).

[SensEmbed]. We finally included the sense-based embeddings proposed in (Iacobacci et al., 2015), since our concepts and fillers are fully disambiguated and linked to WordNet synsets.

5.3. Results

Results are shown in Table 8. The model based on semagrams and enriched with simple tf-idf scores clearly outclasses the others for all the tested values of k. However, similar values are obtained with the semagram-based binary approach. The performance is measured with precision@k scores, i.e., the number of concepts of correct category divided by k. Note that our case is a special one where precision is equal to recall, since the models are evaluated on the bounded set of 300 input concepts.

Model	k**=1**	**=5**	**=10**	**=20**	**=30**
C-Corpus	0.72	0.58	0.49	0.36	0.28
S-Corpus	0.69	0.49	0.35	0.23	0.18
S-Binary	0.90	0.73	**0.60**	0.42	**0.32**
S-Mixture	**0.91**	**0.74**	**0.60**	**0.43**	**0.32**
GloVe	0.67	0.55	0.45	0.33	0.26
fastText	0.72	0.59	0.49	0.35	0.27
ConceptNet	0.79	0.63	0.52	0.39	0.30
SensEmbed	0.66	0.57	0.49	0.36	0.28

Table 8: Precision@k of the models under evaluation.

5.4. Interpretability

Figure 1 shows the top-20 most similar concepts within the category *Artifacts* according to the best performing semagram *S-Mixture* model over the others. Note that the different embedding spaces often show divergent scores, due to their nature and the utilized resources. Instead, with semagrams, each similarity score derives from specific fine-grained semantic units. For example, consider the pair (*jewelry*, *pendant*), which shows high divergence among the models. While this is difficult to interpret or manage with embeddings, semagrams provide supporting semantically-typed and disambiguated information which can be further processed (e.g., Table 9 reports three example pairs of Figure 1). Another important aspect is that, while the existing embeddings give similarities on a paradigmatic basis, our semagram-based vectors also integrate syntagmatic relations.

Concept pairs	**Shared (slot, filler) pairs**
uniform, vest	*generalization*: habiliment, covering, clothing, vesture, wear
parasol, umbrella	*generalization*: protective cover, protection; *shape*: round, circular; *material*: plastic
photocopier, stationery	*generalization*: tool, utensil; *material*: plastic, metal; *user*: secretary

Table 9: Semagram-based semantic similarity grounding of three concept pairs of Figure 1.

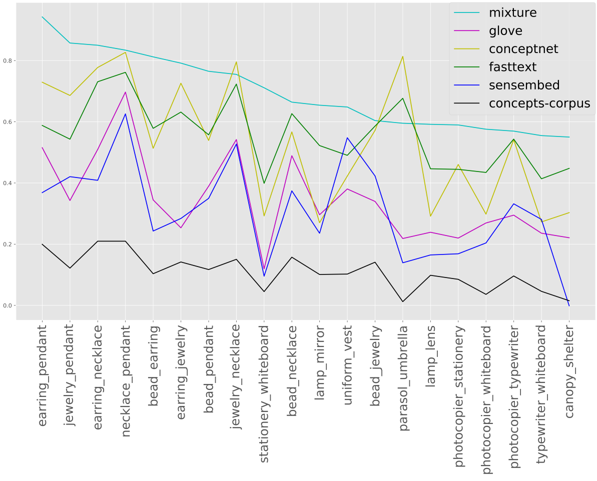

Figure 1: Top-20 most similar concept pairs belonging to the category *Artifacts*, according to *Semagram-Mixture*, in comparison with the other models.

6. Conclusion

In this paper, we started from the notion of *semagram*, i.e., a slot-filler structure to encode word meanings, and proposed a methodology for the creation of a systematized knowledge model of 26 slots integrating and unifying semantic features from different resources. The result of a manual annotation is a semagram base of 50 concepts covering 10 categories, successively extended to 300 concepts through a semi-automated process requiring 1/30 of the original annotation time.

Then, we first showed that an Open Information Extraction approach coupled with a pattern learning method based on WordNet supersenses can be used to extend the semagram base, automatically identifying 4,205 semagrams with a total of 55,245 (slot, filler) pairs with good accuracy levels. Following this we applied an automatic hyponyms-oriented semantic propagation of (slot, filler) pairs through the WordNet taxonomy, reaching high accuracy on a manually-validated test set. Finally, we demonstrated the ability of the model to capture better (and explainable) semantic similarity relations compared to state-of-the-art word embeddings.

As future work, we will integrate the knowledge model with slots for abstract concepts, and extend the knowledge acquisition process by means of crowdsourcing (Poesio et al., 2017) and games-with-a-purpose approaches (Venhuizen et al., 2013; Jurgens and Navigli, 2014). We will also consider the semantic combinations of (Maru et al., 2019)

We release the complete semagram base with the annotation guidelines and the validation data at `http://nlp.uniroma1.it/semagrams`.

Acknowledgments

The last author gratefully acknowledges the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

7. Bibliographical References

Almuhareb, A. and Poesio, M. (2004). Attribute-based and value-based clustering: An evaluation. In *EMNLP*, volume 4, pages 158–165.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proc. of ACL*, pages 86–90. Association for Computational Linguistics.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676.

Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive science*, 34(2):222–254.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Boella, G. and Di Caro, L. (2013). Extracting definitions and hypernym relations relying on syntactic dependencies and support vector machines. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 532–537.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., H. Jr, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3.

Del Corro, L. and Gemulla, R. (2013). Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366. ACM.

Delli Bovi, C., Telesca, L., and Navigli, R. (2015). Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3:529–543.

Devereux, B. J., Tyler, L. K., Geertzen, J., and Randall, B. (2014). The cslb concept property norms. *Behavior research methods*, 46(4):1119–1127.

Di Fabio, A., Conia, S., and Navigli, R. (2019). VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proc. of EMNLP*, pages 1535–1545. Association for Comp. Linguistics.

Fillmore, C. J. (1977). Scenes-and-frames semantics. *Linguistic structures processing*, 59:55–88.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Hanks, P. (2004). Corpus pattern analysis. In *Euralex Proceedings*, volume 1, pages 87–98.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Mit Press.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hovy, E. H., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artif. Intell.*, 194:2–27.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proc. of ACL*, pages 873–882.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015). SensEmbed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105.

Jurgens, D. and Navigli, R. (2014). It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *Trans. of the ACL*, 2:449–464.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., and Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *LREC*, pages 1989–1993. Citeseer.

Maru, M., Scozzafava, F., Martelli, F., and Navigli, R. (2019). SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3532–3538.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behav. r. m.*, 37(4):547–559.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mishra, B. D., Tandon, N., and Clark, P. (2017). Domain-targeted, high precision knowledge extraction. *Transactions of the Association for Computational Linguistics*, 5:233–246.

Moerdijk, F., Tiberius, C., and Niestadt, J. (2008). Accessing the anw dictionary. In *Proc. of the workshop on Cognitive Aspects of the Lexicon*, pages 18–24.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In *Proc. of ACL*, pages 216–225. Association for Computational Linguistics.

Navigli, R. and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In Jan Hajic, et al., editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), July 11-16, 2010, Uppsala, Sweden*, pages 1318–1327. The Association for Computer Linguistics.

Palmer, M., Dang, H. T., and Fellbaum, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Nat.Lan.Eng.*, 13(02):137–163.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Poesio, M., Chamberlain, J., and Kruschwitz, U. (2017). Crowdsourcing. In *Handbook of Linguistic Annotation*, pages 277–295. Springer.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, 17(4):409–441.

Raganato, A., Bovi, C. D., and Navigli, R. (2016). Automatic construction and evaluation of a large semantically enriched wikipedia. In *IJCAI*, pages 2894–2900.

Scarlini, B., Pasini, T., and Navigli, R. (2020). SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of the 34th Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Schuler, K. K. (2005). Verbnet: A broad-coverage, comprehensive verb lexicon.

Silberer, C., Ferrari, V., and Lapata, M. (2013). Models of semantic representation with visual attributes. In *ACL (1)*, pages 572–582.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in ConceptNet 5. In *LREC*, pages 3679–3686.

Speer, R., Chin, J., and Havasi, C. (2016). Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv:1612.03975*.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Venhuizen, N., Evang, K., Basile, V., and Bos, J. (2013). Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

Wilks, Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial Int.*, 6(1):53–74.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics.