Measuring frame relatedness

Marco Pennacchiotti
Yahoo! Inc.
Santa Clara, CA 95054
pennac@yahoo-inc.com

Michael Wirth
Computational Linguistics
Saarland University, Germany
miwirth@coli.uni-sb.de

Abstract

In this paper we introduce the notion of "frame relatedness", i.e. relatedness among prototypical situations as represented in the FrameNet database. We first demonstrate the cognitive plausibility of that notion through an annotation experiment, and then propose different types of computational measures to automatically assess relatedness. Results show that our measures provide good performance on the task of ranking pairs of frames.

1 Introduction

Measuring relatedness among linguistic entities is a crucial topic in NLP. Automatically assessing the degree of similarity or relatedness between two words or two expressions, is of great help in a variety of tasks, such as Question Answering, Recognizing Textual Entailment (RTE), Information Extraction and discourse processing. Since the very beginning of computational linguistics, many studies have been devoted to the definition and the implementation of automatic measures for word relatedness (e.g. (Rubenstein and Goodenough, 1965; Resnik, 1995; Lin, 1998; Budanitsky and Hirst, 2006; Mohammad and Hirst, 2006)). More recently, relatedness between lexical-syntactic patterns has also been studied (Lin and Pantel, 2001; Szpektor et al., 2004), to support advanced tasks such as paraphrasing and RTE. Unfortunately, no attention has been paid so far to the definition of relatedness at the more abstract *situational level* – i.e. relatedness between two prototypical actions, events or state-of-affairs, taken out of context (e.g. the situations of *Killing* and *Death*). A prominent definition of "prototypical situation" is given in frame semantics (Fillmore, 1985), where a situation is modelled as a conceptual structure (a *frame*) constituted by the predicates that can evoke the situation, and the semantic roles expressing the situation's participants.

As measures of word relatedness help in discovering if two word occurrences express related concepts, so measures of **frame relatedness** should help to discover if two large text fragments are related or talk about similar situations. Such measures would be valuable in many tasks. For example, consider the following fragment, in the context of discourse processing:

"In the 1950s the Shah initiated Iran 's nuclear research program and developed an ambitious plan to produce 23,000MW from nuclear power. The program was stopped by the Islamic Revolution in 1979, but it was revived later in the decade, when strategic interests began to drive the nuclear program."

The underlined words evoke highly related frames, namely ACTIVITY_START, ACTIVITY_STOP and CAUSE_TO_RESUME. This could suggest to link the three textual fragments associated to the words, into a single coherent discourse unit, where the semantic roles of the different fragments can be easily mapped as co-referential (e.g. "Iran's nuclear research program" - "The program" - "it"). Frame relatedness can also help in RTE. Consider for example the following entailment pair:

Text : "An avalanche has struck a popular skiing resort in Austria, killing at least 11 people."

Hypothesis : "Humans died in an avalanche."

The frames KILLING and DEATH, respectively evoked by *killing* and *died*, are highly related and can then be mapped. Leveraging this mapping, an RTE system could easily discover that the Text entails the Hypothesis, by verifying that the fillers of the mapped semantic roles of the two frames are semantically equivalent.

In this paper we investigate the notion of relatedness in the context of frame semantics, and propose different types of automatic measures to compute relatedness between frames. Our main contributions can be summarized as follows: (1) We empirically show that the notion of frame relatedness is intuitive and principled from a cognitive perspective: to support this claim, we report agreement results over a pool of human annotators on the task of ranking frame pairs on relatedness; (2) We propose a variety of measures for computing frame relatedness, inspired by different approaches and by existing measures for word relatedness; (3) We show that our measures offer good performance, thus opening the path to the use of frame relatedness as a practical tool for NLP, and showing that measures for word relatedness can be successfully adapted to frames. The paper is organized as follows. In Section 2 we summarize related work. In Section 3 we describe the experiment of humans ranking frame pairs, and discuss the results. In Section 4 and 5 we respectively introduce our relatedness measures, and test them over a manual gold standard. In Section 6 we draw final conclusions and outline future work.

2 Related Work

Much research in NLP has studied similarity and relatedness between words. Rubenstein and Goodenough (1965) were the first to propose a procedure to assess human agreement on ranking pairs of words on relatedness. Their experiment was later replicated by Resnik (1995) and Charles (2000). All these studies reported good levels of agreements among annotators, suggesting that the notion of word relatedness is cognitively principled. In our experiment in Section 3.2 we apply the same procedure to assess agreement on ranking frames.

Measures for estimating word relatedness have been systematically proposed since the early 90's, and are today widely used in NLP for various tasks. Most measures can be classified either as corpus-based or ontology-based. *Corpus-based measures* compute relatedness looking at the distributional properties of the two words: words that tend to co-occur in the same contexts or having similar distributional profiles, are deemed to be highly related. A complete survey on these measures is reported in (Mohammad and Hirst, 2006). *Ontology-based measures* estimate relatedness by studying the path connecting the two words in an ontology or a hierarchical lexicon (e.g. WordNet). The basic idea is that closer words are more related than distant ones. Budanitsky and Hirst (2006) provide an extensive survey of these measures.

Budanitsky and Hirst (2006) also point out an important distinction, between *relatedness* and *similarity*. Two words are related if any type of relation stands between them, e.g. antonymy or meronymy; they are similar when related through an *is-a* like hierarchy. Similarity is then a special case of relatedness. Following Budanitsky and Hirst (2006), we consider two frames as *similar* if they are linked via *is-a* like relations (e.g. GETTING and COMMERCE_BUY), while as *related* if any relation stands between them (e.g. causation between KILLING and DEATH). In this paper, we focus our attention solely on the notion of frame relatedness.

3 Defining frame relatedness

In this section we check if the notion of frame relatedness is intuitive and principled from a cognitive perspective. In Section 3.1 we first introduce the basic concepts or frame semantics; in Section 3.2 we report the agreement results obtained by human annotators, on the task of ranking a dataset of frame pairs according to relatedness.

3.1 Frame Semantics and FrameNet

Frame semantics (Fillmore, 1985) seeks to describe the meaning of a sentence as it is actually understood by characterizing the background knowledge necessary to understand the sentence. Background knowledge is represented in the form of *frames*, conceptual structures modelling prototypical situations. Linguistically, a frame is a semantic class containing predicates called *lexical units* (LU), that can *evoke* the described situation (see example in Table 1). Each frame comes with its own set of semantic roles, called *frame elements* (FE). These are the participants and props in the abstract situation described. Roles are local to individual frames, thus avoiding the commitment to a small set of universal roles, whose specification has turned out to be unfeasible in the past.

The Berkeley FrameNet project (Baker et al., 1998) has been developing a frame-semantic lexicon for the core vocabulary of English since 1997. The current FrameNet release contains about 800 frames and 10,000 lexical units. Part of FrameNet

Frame: STATEMENT		
This frame contains verbs and nouns that communicate the act of a SPEAKER to address a MESSAGE to some ADDRESSEE using language. A number of the words can be used performatively, such as *declare* and *insist.*		
FEs	SPEAKER	**Evelyn** said she wanted to leave.
	MESSAGE	Evelyn announced **that she wanted to leave**.
	ADDRESSEE	Evelyn spoke **to me** about her past.
	TOPIC	Evelyn's statement **about her past**
	MEDIUM	Evelyn preached to me **over the phone**.
LUs	acknowledge.v, acknowledgment.n, add.v, address.v, admission.n, admit.v, affirm.v, affirmation.n, allegation.n, allege.v, announce.v, ...	

Table 1: Example frame from FrameNet.

is also a corpus of annotated example sentences from the British National Corpus, currently containing 135,000 sentences.

In FrameNet, asymmetric frame relations can relate two frames, forming a complex hierarchy (Ruppenhofer et al., 2005): *Inheritance*: anything true in the semantics of the parent frame, must also be true for the other (e.g. KILLING – EXECUTION). *Uses*: a part of the situation evoked by one frame refers to the other. *Subframe*: one frame describes a subpart of a complex situation described in the other (e.g. CRIMINAL-PROCESS – SENTENCING). *Causative_of* : the action in one frame causes the event described in the other (e.g. KILLING – DEATH). *Inchoative_of*: the event in one frame ends in the state described in the other (e.g. DEATH – DEAD_OR_ALIVE). *Precedes*: one frame temporally proceeds the other (e.g. FALL_ASLEEP – SLEEP). *Perspective_on*: one frame describes a specific point-of-view on a neutral frame.

The first two are *is-a* like relations, while the others are non-hierarchical.

3.2 Manually ranking related frames

We asked a pool of human annotators to manually rank a set of frame pairs according to their relatedness. The goal was twofolds. First, we wanted to check how intuitive the notion of frame relatedness is, by computing inter-annotator agreement, and by comparing the agreement results to those obtained by Rubenstein and Goodenough (1965) for word relatedness. Second, we planned to use the produced dataset as a gold standard for testing the relatedness measures, as described in Section 5. In the rest of the section we describe the annotation process in detail.

Dataset creation. We created two different datasets, a *simple* and a *controlled set*, each containing 155 pairs. Frame pairs in the *simple set* were randomly selected from the FrameNet database. Frame pairs in the *controlled set* were either composed of two frames belonging to the same scenario[1], or being so that one frame is one edge from the scenario of the other. This ensured that all pairs in the controlled set contained semantically related frames. Indeed, we use the controlled set to check if human agreement and automatic measure accuracy get better when considering only highly related frames.

Human ranking agreement. A preliminary annotation phase involved a group of 15 annotators consisting of graduate students and researchers, native or nearly native speakers of English. For each set, each annotator was given 15 frame pairs from the original 155 set: 5 of these where shared with all other annotators. This setting has three advantages: (1) The set is small enough to obtain a reliable annotation in a short time; (2) We can compute the agreement among the 15 annotators over the shared pairs; (3) We can check the reliability of the final gold standard created in the second phase (see following section) by comparing to the annotations. Each annotator was asked to order a shuffled deck of 15 cards, each one describing a pair of frames. The card contained the following information about the two frames: names; definitions; the lists of core FEs; a frame annotated sentence for each frame, randomly chosen from the FrameNet database. Similarly to Rubenstein and Goodenough (1965) we gave the annotators the following instructions: (i) After looking through the whole deck, order the pairs according to amount of relatedness; (ii) You may assign the same rank to pairs having the same degree of relatedness (i.e. ties are allowed).

We checked the agreement among the 15 annotators in ranking the 5 shared pairs by using the Kendall's τ correlation coefficient (Kendall, 1938). Kendall's τ can be interpreted as the difference between the probability that in the dataset two variables are in the same order versus the probability that they are in different orders (see (Lapata, 2006) for details). The average corre-

[1] A scenario frame is a "hub" frame describing a general topic; specific frames modelling situations related to the topic are linked to it (e.g. COMMERCE_BUY and COMMERCIAL_TRANSACTION are linked to COMMERCE_SCENARIO). FrameNet contains 16 scenarios.

lation[2] among annotators on the simple and controlled sets was $\tau = 0.600$ and $\tau = 0.547$.

Gold standard ranking. The final dataset was created by two expert annotators, jointly working to rank the 155 pairs collected in the data creation phase. We computed the rank correlation agreement between this annotation and the 15 annotation produced in the first stage. We obtained an average Kendall's $\tau = 0.530$ and $\tau = 0.566$ respectively on the simple and controlled sets (Standard deviations from the average are $StdDev = 0.146$ and $StdDev = 0.173$). These results are all statistically significant at the 99% level, indicating that the notion of "frame relatedness" is intuitive and principled for humans, and that the final datasets are reliable enough to be used as gold standard for our experiments. Table 2 reports the first and last 5 ranked frame pairs for the two datasets.

We compared the correlation results obtained above on "frame relatedness", to those derived from previous works on "word relatedness". This comparison should indicate if ranking related frames (i.e. situations) is more or less complex and intuitive than ranking words.[3] As for words, we computed the average Kendall's τ among three different annotation efforts (namely, (Rubenstein and Goodenough, 1965; Resnik, 1995; Charles, 2000)) carried out over a same dataset of 28 word pairs originally created by Rubenstein and Goodenough. Note that the annotation schema followed in the three works is the same as ours. We obtained a Kendall's $\tau = 0.775$, which is statistically significant at the 99% level. As expected, the correlation for word relatedness is higher than for frames: Humans find it easier to compare two words than two complex situations, as the former are less complex linguistic entities than the latter.

4 Measures for frame relatedness

Manually computing relatedness between all possible frame pairs in FrameNet is an unfeasible task. The on-going FrameNet project and automatic methods for FrameNet expansion (e.g. (Pennacchiotti et al., 2008)) are expected to produce an ever growing set of frames. The definition of automatic measures for frame relatedness is thus a key issue. In this section we propose different types of such measures.

4.1 WordNet-based measures

WordNet-based measures estimate relatedness by leveraging the WordNet hierarchy. The hypothesis is that two frames whose sets of LUs are close in WordNet are likely to be related. We assume that LUs are sense-tagged, i.e. we know which WordNet senses of a LU map to a given frame. For example, among the 25 senses of the LU *charge.v*, only the sense *charge.v#3* ("demand payment") maps to the frame COMMERCE_COLLECT.

Given a frame F, we define S_F as the set of all WordNet senses that map to any frame's LU (e.g. for COMMERCE_COLLECT, S_F contains *charge.v#3*, *collect.v#4*, *bill.v#1*). A generic WordNet-based measure is then defined as follows:

$$wn(F_1, F_2) = \frac{\sum_{s1 \in S_{F_1}} \sum_{s2 \in S_{F_2}} wn\_rel(s1, s2)}{|S_{F_1}| \cdot |S_{F_2}|} \quad (1)$$

where $wn\_rel(s1, s2)$ is a sense function estimating the relatedness between two senses in WordNet. Since we focus on frame relatedness, we are interested in assigning high scores to pairs of senses which are related by any type of relations in WordNet (i.e. not limited to *is-a*). We therefore adopt as function wn_rel the Hirst-St.Onge measure (Hirst and St.Onge, 1998) as it accounts for different relations. We also experiment with the Jiang and Conrath's (Jiang and Conrath, 1997) measure which relies only on the *is-a* hierarchy, but proved to be the best WordNet-based measure in the task of ranking words (Budanitsky and Hirst, 2006). We call the frame relatedness measures using the two functions respectively as $wn\_hso(F_1, F_2)$ and $wn\_jcn(F_1, F_2)$.

4.2 Corpus-based measures

Corpus-based measures compute relatedness looking at the distributional properties of the two frames over a corpus. The intuition is that related frames should occur in the same or similar contexts.

[2] Average correlation is computed by averaging the τ obtained on each pair of annotators, as suggested in (Siegel and Castellan, 1988); note that the obtained value corresponds to the Kendall u correlation coefficient. Ties are properly treated with the correction factor described in (Siegel and Castellan, 1988).

[3] The comparison should be taken only as indicative, as words can be ambiguous while frames are not. A more principled comparison should involve word senses, not words.

SIMPLE SET	CONTROLLED SET
Measure volume - Measure mass (1)	Knot creation - Rope manipulation (1,5)
Communication manner - Statement (2)	Shoot projectiles - Use firearm (1,5)
Giving - Sent items (3)	Scouring - Scrutiny (3)
Abundance - Measure linear extent (4)	Ambient temperature - Temperature (4)
Remembering information - Reporting (5)	Fleeing - Escaping (5)
...	...
Research - Immobilization (126)	Reason - Taking time (142)
Resurrection - Strictness (126)	Rejuvenation - Physical artworks (142)
Social event - Word relations (126)	Revenge - Bungling (142)
Social event - Rope manipulation (126)	Security - Likelihood (142)
Sole instance - Chatting (126)	Sidereal appearance - Aggregate (142)

Table 2: Human gold standard ranking: first and last 5 ranked pairs (in brackets ranks allowing ties).

4.2.1 Co-occurrence measures

Given two frames F_1 and F_2, the **co-occurrence measure** computes relatedness as the pointwise mutual information (pmi) between them:

$$pmi(F_1, F_2) = log_2 \frac{P(F_1, F_2)}{P(F_1)P(F_2)} \quad (2)$$

Given a corpus C consisting of a set of documents $c \in C$, we estimate pmi as the number of contexts in the corpus (either documents or sentences)[4] in which the two frames co-occur:

$$cr\_occ(F_1, F_2) = log_2 \frac{|C_{F_1,F_2}|}{|C_{F_1}||C_{F_2}|} \quad (3)$$

where C_{F_i} is the set of documents in which F_i occurs, and C_{F_1,F_2} is the set of documents in which F_1 and F_2 co-occur. A frame F_i is said to occur in a document if at least one of its LUs l_{F_i} occurs in the document, i.e.:

$$C_{F_i} = \{c \in C : \exists l_{F_i} \text{ in } c\} \quad (4)$$

$$C_{F_1,F_2} = \{c \in C : \exists l_{F_1} \text{ and } \exists l_{F_2} \text{ in } c\} \quad (5)$$

A limitation of the above measure is that it does not treat ambiguity. If a word is a LU of a frame F , but it occurs in a document with a sense $s \notin S_F$, it still counts as a frame occurrence. For example, consider the word *charge.v*, whose third sense *charge.v#3* maps in FrameNet to COMMERCE_COLLECT. In the sentence: *"Tripp Isenhour was charged with killing a hawk on purpose"*, *charge.v* co-occurs with *kill.v*, which in FrameNet maps to KILLING. The sentence would then result as a co-occurrence of the two above frames. Unfortunately this is not the case, as the sentence's sense *charge.v#2* does not map to the frame. Ideally, one could solve the problem by using a sense-tagged corpus where senses' occurrences are mapped to frames. While sense-to-frame mappings exist (e.g. mapping between frames and WordNet senses in (Shi and Mihalcea, 2005)), sense-tagged corpora large enough for distributional studies are not yet available (e.g., the SemCor WordNet-tagged corpus (Miller et al., 1993) consists of only 700,000 words).

We therefore circumvent the problem, by implementing pmi in a **weighted co-occurrence measure**, which gives lower weights to co-occurrences of ambiguous words:

$$cr\_wgt(F_1, F_2) = log_2 \frac{\sum_{c \in C_{F_1,F_2}} w_{F_1}(c) \cdot w_{F_2}(c)}{\sum_{c \in C_{F_1}} w_{F_1}(c) \cdot \sum_{c \in C_{F_2}} w_{F_2}(c)} \quad (6)$$

The weighting function $w_F(c)$ estimates the probability that the document c contains a LU of the frame F in the correct sense. Formally, given the set of senses S_l of a LU (e.g. *charge.v#1...charge.v#24*), we define S_{l_F} as the set of senses mapping to the frame (e.g. *charge.v#3* for COMMERCE_COLLECT). The weighting function is then:

$$w_F(c) = \underset{l_F \in L_F \text{ in } c}{\arg\max} \; P(S_{l_F}|l_F) \quad (7)$$

where L_F is the set of LUs of F. We estimate $P(S_{l_F}|l_F)$ by counting sense occurrences of l_F over the SemCor corpus:

$$P(S_{l_F}|l_F) = \frac{|S_{l_F}|}{|S_l|} \quad (8)$$

In other terms, a frame receives a high weight in a document when the document contains a LU whose most frequent senses are those mapped to the frame.[5] For example, in the sentence: *"Tripp Isenhour was charged with killing a hawk on purpose."*, $w_F(c) = 0.17$, as *charge.v#3* is not very frequent in SemCor.

[4] For sake of simplicity in the rest of the section we refer to documents, but the same holds for sentences.

[5] In Eq.8 we use Lidstone smoothing (Lidstone, 1920) to account for unseen senses in SemCor. Also, if a LU does not occur in SemCor, an equal probability (corresponding to the inverse of the number of word's senses) is given to all senses.

4.2.2 Distributional measure

The previous measures promote (i.e. give a higher rank to) frames co-occurring in the *same* contexts. The distributional measure promotes frames occurring in *similar* contexts. The distributional hypothesis (Harris, 1964) has been widely and successfully used in NLP to compute relatedness among words (Lin, 1998), lexical patterns (Lin and Pantel, 2001), and other entities. The underlying intuition is that target entities occurring in similar contexts are likely to be semantically related. In our setting, we consider either documents and sentences as valid contexts.

Each frame F is modelled by a distributional vector $\vec{F}$, whose dimensions are documents. The value of each dimension expresses the association ratio $A(F, c)$ between a document c and the frame. We say that a document is highly associated to a frame when most of the FrameNet LUs it contains, map to the given frame in the correct senses:

$$A(F,c) = \frac{\sum_{l \in L_F \text{ in } c} P(S_{l_F}|l_F)}{\sum_{F_i \in \mathcal{F}} \sum_{l \in F_i \text{ in } c} P(S_{l_{F_i}}|l_{F_i})} \quad (9)$$

where $\mathcal{F}$ is the set of all FrameNet frames, and $P(S_{l_F}|l_F)$ is as in Eq. 8. We then compute relatedness between two frames using cosine similarity:

$$cr\_dist(F_1, F_2) = \frac{\vec{F_1} \cdot \vec{F_2}}{|\vec{F_1}| * |\vec{F_2}|} \quad (10)$$

When we use sentences as contexts we refer to $cr\_dist\_sent(F_1, F_2)$, otherwise to $cr\_dist\_doc(F_1, F_2)$

4.3 Hierarchy-based measures

A third family or relatedness measures leverages the FrameNet hierarchy. The hierarchy forms a directed graph of 795 nodes (frames), 1136 edges, 86 roots, 7 islands and 26 independent components. Similarly to measures for word relatedness, we here compute frame relatedness leveraging graph-based measures over the FrameNet hierarchy. The intuition is that the closer in the hierarchy two frames are, the more related they are[6]. We here experiment with the Hirst-St.Onge and the Wu and Palmer (Wu and Palmer, 1994) measures, as they are pure taxonomic measures, i.e. they do not require any corpus statistics.

[6]The *Pathfinder Through FrameNet* tool gives a practical proof of this intuition: `http://fnps.coli.uni-saarland.de/pathsearch`.

WU and Palmer: this measure calculates relatedness by considering the depths of the two frames in the hierarchy, along with the depth of their least common subsumer (LCS):

$$hr\_wu(F_1, F_2) = \frac{2 \cdot dp(LCS)}{ln(F_1, LCS) + ln(F_2, LCS) + 2 \cdot dp(LCS)} \quad (11)$$

where ln is the length of the path connecting two frames, and dp is the length of the path between a frame and a root. If a path does not exist, then $hr\_wu(F_1, F_2) = 0$.

Hirst-St.Onge: two frames are semantically close if they are connected in the FrameNet hierarchy through a "not too long path which does not change direction too often":

$$hr\_hso(F_1, F_2) = M - \text{ path length } - k \cdot d \quad (12)$$

where M and and k are constants, and d is the number of changes of direction in the path. If a path does not exist, $hr\_hso(F_1, F_2) = 0$. For both measures we consider as valid edges all relations.

The FrameNet hierarchy also provides for each relation a partial or complete FE mapping between the two linked frames (for example the role *Victim* of KILLING maps to the role *Protagonist* of DEATH). We leverage this property implementing a **FE overlap measure**, which given the set of FEs of the two frames, FE_1 and FE_2 , computes relatedness as the percentage of mapped FEs:

$$hr\_fe(F_1, F_2) = \frac{|FE_1 \cap FE_2|}{max(|FE_1|, |FE_2|)} \quad (13)$$

The intuition is that FE overlap between frames is a more fine grained and accurate predictor of relatedness wrt. simple frame relation measures as those above – i.e. two frames are highly related not only if they describe connected situations, but also if they share many participants.

5 Experiments

We evaluate the relatedness measures by comparing their rankings over the two datasets described in Section 3.2, using the manual gold standard annotation as reference. As evaluation metrics we use Kendall's τ. As baselines, we adopt a *definition overlap measure* that counts the percentage of overlapping content words in the definition of the two frames;[7] and a *LU overlap baseline*

[7]We use stems of nouns, verbs and adjectives.

Measure	*Simple Set*	*Controlled Set*
wn_jcn	0.114	0.141
wn_hso	0.106	0.141
cr_occ_sent	0.239	0.340
cr_wgt_sent	**0.281**	**0.349**
cr_occ_doc	0.143	0.227
cr_wgt_doc	0.173	0.240
cr_dist_doc	0.152	0.240
hr_wu	0.139	0.286
hr_hso	0.134	0.296
hr_fe	0.252	0.326
def overlap baseline	*0.056*	*0.210*
LU overlap baseline	*0.080*	*0.253*
human upper bound	*0.530*	*0.566*

Table 3: Kendall's τ correlation results for different measures over the two dataset.

that counts the percentage of overlapping LUs between the two frames. We also defined as *upper-bound* the human agreement over the gold standard. As regards distributional measures, statistics are drawn from the TREC-2002 Vol.2 corpus, consisting of about 110 million words, organized in 230,401 news documents and 5,433,048 sentences[8]. LUs probabilities in Eq. 8 are estimate over the SemCor 2.0 corpus, consisting of 700,000 running words, sense-tagged with WordNet 2.0 senses.[9]. WordNet-based measures are computed using WordNet 2.0 and implemented as in (Patwardhan et al., 2003). Mappings between WordNet senses and FrameNet verbal LUs are taken from Shi and Mihalcea (2005); as mappings for nouns and adjectives are not available, for the WordNet-based measures we use the first sense heuristic.

Note that some of the measures we adopt need some degree of supervision. The WordNet-based and the *cr_wgt* measures rely on a WordNet-FrameNet mapping, which has to be created manually or by some reliable automatic technique. Hierarchy-based measures instead rely on the FrameNet hierarchy that is also a manual artifact.

5.1 Experimental Results

Table 3 reports the correlation results over the two datasets. Table 4 reports the best 10 ranks produced by some of the best performing measures. Results show that all measures are positively correlated with the human gold standard, with a level of significance beyond the $p < 0.01$ level , but the wn_jcn measure which is at $p < 0.05$. All measures, but the WordNet-based ones, significantly outperform the *definition overlap* baseline on both datasets, and most of them also beat the more informed *LU overlap* baseline.[10] It is interesting to notice that the two best performing measures, namely *cr_wgt_sent* and *hr_fe*, use respectively a distributional and a hierarchy-based strategy, suggesting that both approaches are valuable. WordNet-based measures are less effective, performing close or below the baselines.

Results obtained on the simple set are in general lower than those on the controlled set, suggesting that it is easier to discriminate among pairs of connected frames than random ones. A possible explanation is that when frames are connected, all measures can rely on meaningful evidence for most of the pairs, while this is not always the case for random pairs. For example, corpus-based measures tend to suffer the problem of data sparseness much more on the simple set, because many of the pairs are so loosely related that statistical information cannot significantly emerge from the corpus.

WordNet-based measures. The low performance of these measures is mainly due to the fact that they fail to predict relatedness for many pairs, e.g. *wn_hso* assigns zero to 137 and 119 pairs, respectively on the simple and controlled sets. This is mostly caused by the limited set of relations of the WordNet database. Most importantly in our case, WordNet misses the *situational relation* (Hirst and St.Onge, 1998), which typically relates words participating in the same situation (e.g. *child care* - *school*). This is exactly the relation that would help in mapping frames' LUs. Another problem relates to adjectives and adverbs: WordNet measures cannot be trustfully applied to these part-of-speech, as they are not hierarchically organized. Unfortunately, 18% of FrameNet LUs are either adjectives or adverbs, meaning that such amount of useful information is lost. Finally, WordNet has in general an incomplete lexical coverage: Shi and Mihalcea (2005) show that 7% of FrameNet verbal LUs do not have a mapping in WordNet.

Corpus-based measures. Table 3 shows that co-occurrence measures are effective when using

[8]For computational limitations we could not afford experimenting the *cr_dist_sent* measure, as the number and size of the vectors was too big.

[9]We did not use directly the SemCor for drawing distributional statistics, because of its small size.

[10]The average level of correlation obtained by our measures is comparable to that obtained in other complex information-ordering tasks, e.g. measuring compositionality of verb-noun collations (Venkatapathy and Joshi, 2005)

WN_JCN	CR_WGT_SENT	HR_FE
Ambient temperature - Temperature (4)	Change of phase - Cause change of phase (7)	Shoot projectiles - Use firearm (1,5)
Run risk - Endangering (27)	Knot creation - Rope manipulation (1,5)	Intentionally affect - Rope manipulation (37,5)
Run risk - Safe situation (51)	Ambient temperature - Temperature (4)	Knot creation - Rope manipulation (1,5)
Knot creation - Rope manipulation (1,5)	Shoot projectiles - Use firearm (1,5)	Ambient temperature - Temperature (4)
Endangering - Safe situation (62)	Hit target - Use firearm (18)	Hit target - Intentionally affect (91,5)
Shoot projectiles - Use firearm (1,5)	Run risk - Safe situation (51)	Safe situation - Security (28)
Scouring - Scrutiny (3)	Safe situation - Security (28)	Suspicion - Criminal investigation (40)
Reliance - Contingency (109)	Cause impact - Hit target (10)	Age - Speed (113)
Safe situation - Security (28)	Rape - Arson (22)	Motion noise - Motion directional (55)
Change of phase - Cause change of phase (7)	Suspicion - Robbery (98)	Body movement - Motion (45)

Table 4: First 10 ranked frame pairs for different relatedness measure on the Controlled Set; in brackets, the rank in the gold standard (full list available at *(suppressed)*).

sentences as contexts, while correlation decreases by about 10 points using documents as contexts. This suggest that sentences are suitable contextual units to model situational relatedness, while documents (i.e. news) may be so large to include unrelated situations. It is interesting to notice that corpus-based measures promote frame pairs which are in a non-hierarchical relation, more than other measures do. For example the pair CHANGE OF PHASE - CAUSE CHANGE OF PHASE score first, and RAPE - ARSON score ninth, while the other measures tend to rank them much lower. By contrast, the two frames SCOURING - INSPECTING which are siblings in the FrameNet hierarchy and rank 17th in the gold standard, are ranked only 126th by *cr_wgt_sent*. This is due to the fact that hierarchically related frames are substitutional – i.e. they tend not to co-occur in the same documents; while otherwise related frames are mostly in syntagmatic relation. As for *cr_dist_doc*, it performs in line with *cr_wgt_doc*, but their ranks differ; *cr_dist_doc* promotes more hierarchical relations: distributional methods capture both paradigmatically and syntagmatically related entities.

Hierarchy-based measures. As results show, the FrameNet hierarchy is a good indicator of relatedness, especially when considering FE mappings. Hierarchy-based measures promote frame pairs related by diverse relations, with a slight predominance of *is-a* like ones (indeed, the FrameNet hierarchy contains roughly twice as many is-a relations as other ones). These measures are slightly penalized by the low coverage of the FrameNet hierarchy. For example, they assign zero to CHANGE OF PHASE - ALTERED PHASE, as an inchoative link connecting the frames is missing.

Correlation between measures. We computed the Kendall's τ among the experimented measures, to investigate if they model relatedness in different or similar ways. As expected, measures of the same type are highly correlated (e.g. *hr_fe* and *hr_wu* have $\tau = 0.52$), while those of different types seem complementary, showing negative or non-significant correlation (e.g. *cr_wgt_sent* has $\tau = -0.034$ with *hr_wu*, and $\tau = 0.078$ with *wn_jcn*). The *LU overlap baseline* shows significant correlation only with *hr_wu* ($\tau = 0.284$), suggesting that in the FrameNet hierarchy frames correlated by some relation do share LUs.

Comparison to word relatedness. The best performing measures score about 0.200 points below the human upper bound, indicating that ranking frames is much easier for humans than for machines. A direct comparison to the word ranking task, suggests that ranking frames is harder than words, not only for humans (as reported in Section 3.2), but also for machines: Budanitsky and Hirst (2006) show that measures for ranking words get much closer to the human upper-bound than our measures do, confirming that frame relatedness is a fairly complex notion to model.

6 Conclusions

We empirically defined a notion of frame relatedness. Experiments suggest that this notion is cognitively principled, and can be safely used in NLP tasks. We introduced a variety of measures for automatically estimating relatedness. Results show that our measures have good performance, all statistically significant at the 99% level, though improvements are expected by using other evidence. As future work, we will build up and refine these basic measures, and investigate more complex ones. We will also use our measures in applications, to check their effectiveness in supporting various tasks, e.g. in mapping frames across Text and Hypothesis in RTE, in linking related frames in discourse, or in inducing frames for LU which are not in FrameNet (Baker et al., 2007).

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 Task 19: Frame Semantic Structure Extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic, June.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Walter Charles. 2000. Contextual correlates of meaning. *Applied Psycholinguistics*, (21):502–524.

C. J. Fillmore. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica*, IV(2).

Zellig Harris. 1964. Distributional structure. In Jerrold J. Katz and Jerry A. Fodor, editors, *The Philosophy of Linguistics*, New York. Oxford University Press.

Graeme Hirst and David St.Onge, 1998. *Lexical chains as representations of context for the detection and correction of malapropisms*, pages 305–332. MIT press.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics (ROCLING X)*, Taiwan.

Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*, (30):81–93.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

G.J. Lidstone. 1920. Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192.

Dekang Lin and Patrick Pantel. 2001. DIRT-discovery of inference rules from text. In *Proceedings of KDD-01*, San Francisco, CA.

Dekang Lin. 1998. Automatic retrieval and clustering of similar word. In *Proceedings of COLING-ACL*, Montreal, Canada.

G. A. Miller, C. Leacock, T. Randee, and Bunker R. 1993. A Semantic Concordance. In *In Proceedings of the 3rd DARPA Workshop on Human Language Technology*, Plainsboro, New Jersey.

Saif Mohammad and Graeme Hirst. 2006. Distributional measures of concept-distance. a task-oriented evaluation. In *Proceedings of EMNLP-2006*, Sydney,Australia.

S. Patwardhan, S. Banerjee, and T. Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.

Marco Pennacchiotti, Diego De Cao, Paolo Marocco, and Roberto Basili. 2008. Towards a vector space model for framenet-like resources. In *Proceedings of LREC*, Marrakech, Marocco.

Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada.

H. Rubenstein and J.B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson. 2005. FrameNet II: Extended Theory and Practice. In *ICSI Technical Report*.

Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *In Proceedings of Cicling*, Mexico.

S. Siegel and N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Idan Szpektor, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcellona, Spain.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Measuring the relative compositionality of verb noun (V-N) collocations by integrating features. In *Proceedings of HLT/EMNLP*, Vancouver, Canad.

Z. Wu and M. Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico.