Software-Projekt

Forschungsplan

26. November 2020

Denis Logvinenko, Natalia Minakova, Livia Zöbeli

Problemstellung:

Aktuelle Evaluationsalgorithmen (SMatch, S2Match, SemBleu) sind nicht imstande, “graded compositional similarity” in AMRs adäquat zu messen.
Anders formuliert besteht das Problem darin, dass die Ähnlichkeit von AMRs, die aus Paraphrasen generiert werden, wegen 1-1 Tripel-Mapping nicht korrekt erfasst werden kann.

Ziel:

Bessere Evaluation der linguistischen Variation (semantisch äquivalenter Ausdrücke in AMR1 und AMR2) in den AMRs, die aus Text generiert worden sind.

Methode:

In Schritt 1 analysieren wir das Problem anhand bekannter Metriken und schlagen in Schritt 2 einen alternativen Score zum Vergleich vor, wobei die Ergebnisse aus Schritt 1 berücksichtigt werden.

1. Charakterisierung des Problems:

Auswertung der aktuellen Metriken und Ermittlung von bestimmten Stellen, bei denen die aktuellen Metriken nicht gut sind. Zum Beispiel, welche Token / Phrasen ähnlich sind, aber von der Metrik nicht erfasst werden.

- Vergleich verschiedener AMRs:
    - Ist die AMR tatsächlich die Darstellung der Phrasen?
    - Wie ähnlich sind die AMR-Scores zu Text-Scores?
        - MSE-Error berechnen zwischen AMR-Score und Gold-Score, um zu sehen, in welchen Sätzen aktuelle Metriken am meisten Probleme haben
            - Bei Beispielen, wo die AMRs schlecht funktionieren: Welche Kanten sind in welcher Kombination beteiligt? Gibt es ein Muster?
            - Funktioniert unser AMRFScore besser?

- Welche Teilausdrücke im Satz sind gleich oder teilweise gleich? Welche Phrasen stehen in einer Beziehung zueinander?
- Welche Token entsprechen welchem Knoten in der Grafik?
  - Vergleich des Graphen mit der Matrix

Daten für die Analyse: Subset von einem STS-Datenset (siehe Sektion “Daten”).

2. AMRFScore (inspiriert von *BERTScore*):

1. Sätze werden syntaktisch geparst.
2. Man berechnet die Kosinusähnlichkeit zwischen ihnen und sortiert sie dementsprechend, was einem erlaubt, eine geeignete Zuordnung zwischen Phrasen zu finden.
3. Der Zuordnung gemäß wird ein Mapping zwischen M Tripeln in AMR1 und N Tripeln in AMR2 erzeugt, wobei der Ähnlichkeitsscore zwischen entsprechenden syntaktischen Phrasen für M+N Tripel übernommen wird. Beispiel (C steht für “candidate”, R für “Reference”):

$$S1 = \text{”The kitten runs”}, S2 = \text{”A young cat walks towards a mouse”}$$
$$NP1 = [\text{”The kitten”}], NP2 = [\text{”A young cat”}, \text{”a mouse”}]$$
$$\cos(NP1_0, NP2_0) = 0.69$$
$$\cos(NP1_0, NP2_1]) = 0.27 \wedge$$
$$\forall c_x \in XP_C, r_y \in XP_R : \cos(c_x, r_y) = \cos(XP_C, XP_R)$$
$$AMR_R = \text{(r / run-02}$$
$$\text{:ARG0 (k / kitten))} \rightarrow \text{Reference},$$
$$AMR_C = \text{((w / walk-01}$$
$$\text{:ARG0 (c / cat}$$
$$\text{:mod (y / young))}$$
$$\text{:direction (m / mouse))} \rightarrow \text{Candidate}$$
$$\text{AMRPScore}(AMR_R, AMR_C) = \frac{1}{|AMR_C|} \sum_{x_j^c \in XP_C} \max_{x_i^r \in XP_R} \cos(x_i^r, x_j^c)$$
$$\text{AMRPScore}(AMR_R, AMR_C) = \frac{1}{|AMR_R|} \sum_{x_i^r \in XP_R} \max_{x_j^c \in XP_C} \cos(x_i^r, x_j^c)$$
$$\text{AMRFScore} = 2\frac{PScore \cdot RScore}{PScore + RScore}$$

Daten:

Alle vorläufigen Analysen werden auf STS-Datensets durchgeführt, später eventuell auch auf MT, falls die vorgeschlagene Methode sprachagnostisch angewendet werden kann.
Mögliche Datensätze für Paraphrasen:

- Microsoft Research Paraphrase Corpus (Eine Textdatei mit 5800 Satzpaaren, die aus Nachrichtenquellen im Internet extrahiert wurden, mit menschlichen Anmerkungen, die angeben, ob jedes Paar eine semantische Äquivalenzbeziehung erfasst.
- PPDB
- Andere Datensätze, verfügbar hier

Tools:

- für XP-Parsing: TBD, vorläufig *spaCy* oder ein SOTA-Konstituentenparser
- AMR2Text-Parser (falls kein Ausgangstext zur Verfügung steht; syntaktisches Parsing erfolgt auf der Textebene)
- Text2AMR-Parser (wie ähnlich sind AMRs, die aus Paraphrasen generiert wurden?)
- *SentenceTransformers* basiert auf "*Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*" (dient der Ähnlichkeitsanalyse von gefundenen XPs)

Evaluation:

Es wird auf den Benchmarks evaluiert, die in den in der Sektion "Daten" erwähnten Datensets inkludiert sind.