日本語文からの DBpedia RDF トリプル生成

末木顕人　情報・ネットワーク工学専攻 兼岩憲研究室

RDF と DBpedia

RDF トリプル：

主語 s，述語 p，目的語 o の三つ組 (s, p, o).

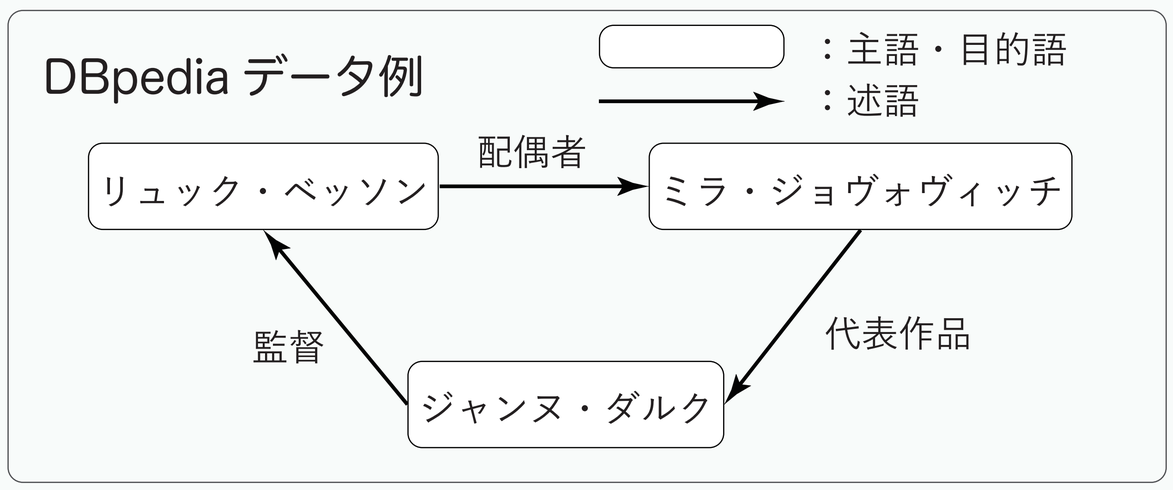
RDF グラフ：

RDF トリプルの集合としてグラフ構造を作る.

DBpedia：

Wikipedia の情報を RDF 形式で表現した知識ベース.
カテゴリツリーや Infobox から半自動的に RDF を生成している.

本文自体は抽出対象ではない



既存手法 [1]

Exner らは英語版 Wikipedia 記事本文から述語項構造を抽出して DBpedia の述語と文中の表現とを対応付け，文章を RDF トリプルに変換する手法を提案した [1].

Besson married actress Milla Jovovich.
主語の項　述語　目的語の項

↓ ・述語項構造解析
・共参照解決

(“Besson” , “marry,” “Milla Jovovich”)

↓ ・Named Entity Linking

(dbr:Luc_Besson, “marry” , dbr:Milla_Jovovich)

↓ ・述語のマッピング
“marry” → dbo:spouse (配偶者)

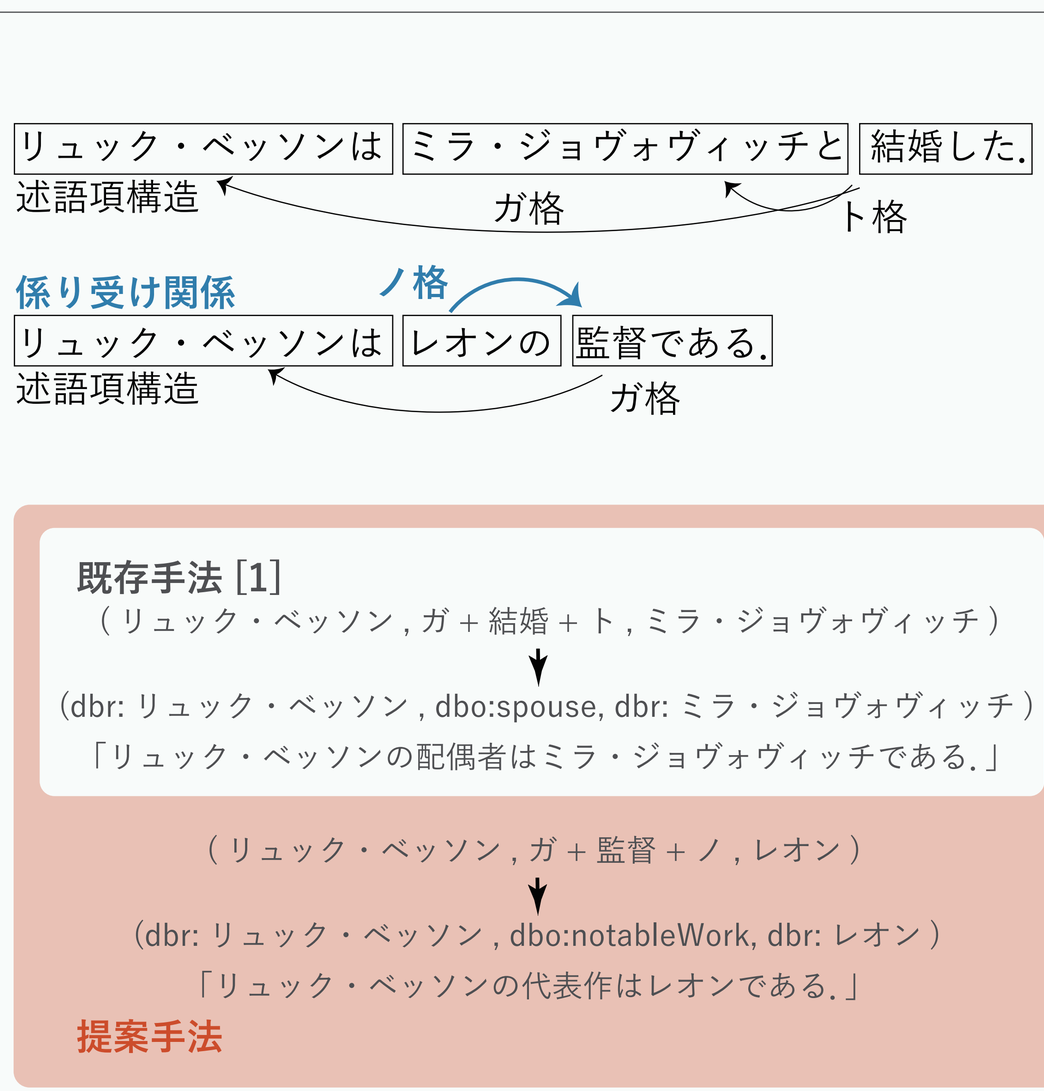
(dbr:Luc_Besson, dbo:spouse, dbr:Milla_Jovovich)

抽出できるのは述語項構造のみ

提案手法

役割関係：

述語項構造に係り受け関係を加えることにより，
「X は Y の R である.」→ $(X, \text{ガ}\ R\ \text{ノ}, Y)$
という関係・トリプルを新たに抽出する.



新たな語間関係により抽出量を増加

実験結果

無作為に選んだ Wikipedia 記事 (97,479 記事) の本文から DBpedia RDF トリプルを抽出した.
既存手法 [1] に対し，生成量を増加・生成精度を向上.

	生成トリプル数	生成トリプル精度
既存手法 [1]	2,881	54.0 %
提案手法	+ 241　3,122	+ 7.0 %　61.0 %

語間関係 (文中の表現)	マッピングした述語
ガ + 受賞 + ヲ	dbo:award
ガ + 子会社 + ノ	dbo:owner
ガ + アナウンサー + ノ	dbo:affiliation

今後の課題

- 構文解析の補完によるトリプル精度の向上
- Wikidata 等，DBpedia 以外のデータへの拡張

参考文献

[1] P. Exner and P. Nugues. Entity Extraction: From Unstructured Text to DBpedia RDF Triples. In Proceedings of the Web of Linked Entities Workshop (WoLE 2012), pages 58–69, 2012.