Tel(s)-Telle(s)-Signs: Highly Accurate Automatic Crosslingual Hypernym Discovery

Ada Wan
University of Zurich
Zurich, Switzerland
ada.wan@uzh.ch

Abstract

We report a highly accurate hypernym discovery heuristic that works on unrestricted texts. This approach leverages morphological cues in French, but given any parallel data and word alignment tool, this proves to be a technique that can work reliably in other languages as well. We tested this method using two French-English corpora of different genres (medical and news) and attained near-perfect accuracy. The key idea is to exploit morphological information in the French trigger phrase *tel*(*s*)*-/telle*(*s*)*- que* (meaning "such as" in English) to uniquely identify the correct hypernym. This shows to be an inexpensive and effective heuristic also when there are multiple noun phrases preceding the trigger phrase, as in the case of prepositional phrase attachment causing ambiguity in interpretation and hypernym acquisition, indicating that this pattern in French is more informative than its English counterpart.

Keywords: hypernym discovery, information extraction, crosslingual knowledge transfer

1. Introduction

The present work focuses on discovering hypernyms in French (FR) using the trigger phrase *tel-*[1] *que* (meaning "such as" in English (EN)), exploiting morphological information to uniquely identify the correct hypernym. This shows to be an inexpensive and effective heuristic also when there are multiple noun phrases preceding the trigger phrase, as in the case of prepositional phrase attachment causing ambiguity in interpretation and hypernym acquisition, proving that this pattern in FR is more informative than its EN counterpart.

In simpler cases, where there is only one noun phrase (NP) preceding the trigger phrase *such as*, hypernym discovery is straightforward. For instance, in the sentence:

[1] *Les agrumes tels que l'orange, le citron ou le pamplemousse contiennent beaucoup de vitamin C.*[2]

and its EN translation:

"Citrus fruits such as oranges, lemons or grapefruits contain a lot of vitamin C."

the NPs preceding *such as* and *tels que* – (*citrus*) *fruits* and *agrumes*, respectively – are the hypernyms denoting the categories or classes in which the hyponyms *oranges*/*orange*, *lemons*/*citron*, and *grapefruits*/*pampelmousse* belong. But seldom does one notice that the bound morpheme *-s* in *agrumes* and *tels* in FR also reveals agreement between the two words and has a role to play in hypernym recognition.

It appears that one could use this morphological information in *tel-* to correctly identify the antecedent of this hyper-/hyponym relation, disambiguating cases where multiple nouns preceding *tel- que* (hereafter: *tq*) pose as plausible hypernym candidates. This would also resolve the issue brought about by the prepositional compound phrase construction, i.e. the combination of a noun phrase and a prepositional phrase (Lefever et al., 2014), alternatively formulated as the error caused by the "local nature of the patterns" (Ritter et al., 2009):

> "A sentence with ... *urban birds in cities such as pigeons* ... matches the pattern 'C such as E' with C bound to *city* and E bound to *pigeon*, leading to *city* as a hypernym of *pigeon*."

(Note that *cities*, the more local noun to *such as*, can be the hypernym in the sentence ... *urban birds in cities such as Paris, Rome* ....)

This paper demonstrates the reliability and accuracy of our proposed agreement pattern in hypernym discovery in FR. Once a hypernym is identified in FR in this manner, reliable extraction of hypernyms in other languages can also be expected given proper word alignment and bitexts.

2. Related Work

Ontological relations are commonly extracted from unstructured texts to build and extend semantic taxonomies or relational databases, such as *WordNet* and *DBpedia*, in addition to industry-specific knowledge bases. Hearst (1992) broke ground with her seminal work in ontological discovery which aimed to automatically acquire hyponyms in unrestricted texts. Both Yamada et al. (2009) and Lefever et al. (2014) noted the myriad of research in this area since then – many approached this topic with pattern-based methods, while others used clustering or distributional similarity ones as well as techniques with word class lattices and embeddings. Yet, except in work by Shinzato & Torisawa (2004), Tjong Kim Sang (2007), Bosma & Vossen (2010), Tjong Kim Sang et al. (2011), Fu et al. (2013), and Lefever et al. (2014), morphology (word-internal structure) has played little role in ontological relation extraction. All of these authors exploited morphology in a similar fashion – they treated the rightmost constituent or affix (roughly considering radicals in Chinese characters as affixes of sorts here) of a complex noun or a multi-word expression as

[1] *tel-* is used as a shorthand for *tel*, *tels*, *telle*, or *telles*, cf. Section 3.1.

[2] `http://bdl.oqlf.gouv.qc.ca/bdl/gabarit\_bdl.asp?id=2437`

the head noun and hence the hypernym. For example, in Japanese: *eiga* as the hypernym of *amerika-eiga* "American movie" and *nihon-eiga* "Japanese movie"; in Dutch: *beleid* "policy" as the hypernym of *landbouwbeleid* "agricultural policy" and *pijpleiding* "pipeline" as the hypernym of *offshore pijpleiding* "offshore pipeline"; and in Chinese: the radical 虫 "insect" as the hypernym of 蜻蜓 "dragonfly" and 企鹅 "penguin" as the hypernym of 皇帝企鹅 "emperor penguin". Lefever et al. also accounted for prepositional compound phrases by designating the head at the left edge of the compound phrase as the hypernym, e.g. *saneren* "remediation" as hypernym of *saneren van verontreinigde bodems* "remediation of contaminated soils". However, this account would not suffice for constructions requiring the more local noun to be the hypernym.

In addition to addressing the issue of how lexico-morphological information can be used in hypernym detection as in the work above, we aim to demonstrate in this paper how leveraging agreement information can help resolve ambiguity in hypernym identification and report how the reliability of these cues can transfer to high accuracy in extracting hypernyms in another language given proper word alignment.

3. Method

3.1. Approach

The discovery of our approach to hypernym recognition was a data-driven endeavor inspired by one lexico-syntactic pattern indicated in Hearst (1992) using *such as*:

NP_0 such as $\{\text{NP}_1,\ \text{NP}_2 \ldots, (\text{and}|\text{or})\}\ \text{NP}_n$

where it is implied that:

for all NP_i, $1 \leq \text{i} \leq \text{n}$, hyponym($\text{NP}_i$, NP_0), i.e. hypernym(NP_0, NP_i).

While examining EN-FR parallel sentences in the EMEA corpus, we learned that *tq* is the most common translation for *such as*. (*que* also has a variant *qu'* – mostly when the next word starts with a vowel.) We noticed that not only does the pattern "NP *tq* NP" manifest a hyper-/hyponym relationship as one would expect in EN, but the word *tel-* also agrees with the preceding noun which is the hypernym of said pattern. Generally, adjectives and certain pronouns in FR agree in gender and number with nouns which they modify or co-refer with. The *tel-* in *tq* is found to vary – in its masculine singular form *tel*, feminine singular *telle*, masculine plural *tels*, and feminine plural *telles*. There is always (at least) one noun preceding *tq* with matching agreement – distant or local. That should not be surprising as there is a rule in FR grammar prescribing that when *tq* introduces a comparison and appears before a single example or an enumeration, *tel-* is to agree with the noun that precedes it, i.e. the one that it exemplifies[3], as can be seen in sentence [1] in Section 1., as well as in [2] below:

[2] *À l'examen, vous pourrez utiliser des livres de référence tels que grammaires et encylopédies.*

(EN translation: "At the examination, you can use reference books such as grammars and encyclopedias.")

tels in [2] agrees with *livres de référence*, more precisely with the head noun *livres*, as they are both masculine plural (mp); it does not agree with other nouns in the sentence, e.g.: *référence* (feminine singular (fs)), *grammaires* (fp), or *encyclopédies* (fp)).

We hypothesize that the closest noun preceding *tq* that agrees with *tel-* in gender and number is the hypernym in the sentence in FR in the pattern[4]:

NP_x... ($\text{NP}_{y \neq x}$)* tel_x que $\{\text{NP}_1, \ldots (\text{et}|\text{ou})\}$ NP_n

where the subscripts x and y indicate gender and number marking, and where:

for all NP_i, $1 \leq \text{i} \leq \text{n}$, hyponym($\text{NP}_i$, NP_x), i.e. hypernym(NP_x, NP_i).

3.2. Experiment

We tested our hypothesis using two sets of FR-EN parallel corpora – the EMEA corpus made from PDF documents from the European MEdicines Agency[5] (Tiedemann, 2009) and the news commentary training data from the WMT 2014 shared task[6]. After tokenization, we filtered sentences with length over 100 words for the EMEA corpus and 200 for the WMT corpus. We used GIZA++ (Och and Ney, 2003) with the grow-diag-final-and heuristic (Koehn et al., 2007) for word alignment, and tagged the FR text using the morphological analyzer *morfette*[7] (Chrupala et al., 2008).

We selected the first 50 unique sentence pairs from each corpus for manual evaluation. For these 100 pairs of sentences, the questions we ask in the evaluation process are:

1. Is the closest noun preceding *tq* that agrees with *tel-* in gender and number the hypernym in the sentence?

2. Did *morfette* predict the correct outcome in question 1 above?

3. Is the hypernym in FR aligned with the hypernym in EN?

A positive result for question 1 would confirm that our hypothesis is accurate, that a hypernym in FR can be predicted using morphological information. A positive result for question 2 would show the hypernym can be easily extracted with the aid of a morphological analyzer like *morfette*. A positive result for question 3 would indicate that hypernyms in EN can also be automatically extracted by

[3] http://bdl.oqlf.gouv.qc.ca/bdl/gabarit_bdl.asp?id=2437

[4] Similar to the corresponding Hearst pattern in EN, our pattern assumes *tq* to be immediately surrounded by Ns/NPs, not verbs (e.g. passive participle in *such as indicated*) or subordinate clauses after *que* (esp. since *que*, like *that/as* in EN, can often be followed by a clausal structure beginning with e.g. a N-V sequence). There should also be no punctuation between *tel-* and its preceding NP.

[5] http://opus.lingfil.uu.se/EMEA.php

[6] http://statmt.org/wmt14/translation-task.html

[7] https://sites.google.com/site/morfetteweb/home

means of this heuristic with word alignment tools such as GIZA++.

4. Results and Discussion

Out of the 100 sentences we evaluated manually, 97 confirm our hypothesis. In 91 of these cases, *morfette* correctly tagged the hypernyms for gender and number, suggesting that the morphological cue can be readily extracted. Finally, crosslingual alignment of the hypernyms succeeded in 87.5[8] instances: the errors here were mostly due to the original sentence alignment being wrong, such that the EN sentences failed to contain the corresponding pattern in the first place. Table 1 provides a summary broken down by corpus.

We noticed that not only can our hypothesis disambiguate prepositional compound phrases but also potentially obviate the notion of headedness or syntactic parsing for the task of hypernym extraction. Below are sets of sample sentences from the EMEA corpus which will help elucidate (gender and number information, as provided by *morfette*, is suffixed here onto FR nouns and *tel-* with an underscore for easier reference: _ms (masculine singular), _fs (feminine singular), _mp (masculine plural), _fp (feminine plural)):

1. Long-distance agreement:

   FR: *APTIVUS, co-administré avec le ritonavir_ms à faible dose_ms, doit être utilisé avec précaution_fs chez les patients_mp pouvant présenter un risque_ms accru de saignement_ms en raison_fs d' un traumatisme_ms, d_ms' une chirurgie_fs ou d' antécédents médicaux autres, ou chez ceux recevant des* **traitements_mp** *connus pour augmenter le risque_ms de saignement_ms* **tels_mp que** *les antiagrégants_mp plaquettaires et les anticoagulants_mp, ou chez ceux qui prennent de la vitamine_fs E.*

   EN: *APTIVUS, co-administered with low dose ritonavir, should be used with caution in patients who may be at risk of increased bleeding from trauma, surgery or other medical conditions, or who are receiving* **medicinal products** *known to increase the risk of bleeding* **such as** *antiplatelet agents and anticoagulants or who are taking supplemental vitamin E.*

   Without knowing the meaning of most of the words in the sentence, we learn that *les antiagrégants plaquettaires* and *les anticoagulants* are hyponyms to *traitements* "medicinal products" in FR because *tels* tells us that the closest preceding noun having the same masculine plural ending is *traitements*, bypassing all nearer neighbors *risque* and *saignement*.

2. Nearest neighbor:

   FR: *Celles- ci incluent l'inhibition_fs de la libération_fs de* **cytokines_fp** *proinflammatoires* **telles_fp que** *IL-4, IL-6, IL-8, et IL-13 par les mastocytes_mp basophiles_fp humains, ainsi que l'inhibition_ms de l'expression_fs de la molécule_fs d'adhésion_fs P-sélectine_fs sur les cellules_mp endothéliales_mp.*

   EN: *These include inhibiting the release of proinflammatory* **cytokines such as** *IL-4, IL-6, IL-8, and IL-13 from human mast cells/ basophils, as well as inhibition of the expression of the adhesion molecule P-selectin on endothelial cells.*

   The first NP in the FR sentence in this sentence pair can be literally translated as "the inhibition of the release of proinflammatory cytokines such as IL-4...". If one had no idea what these hyponyms *IL-4*, *IL-6* etc. could be, one could potentially interpret these as kinds of inhibition or release. The head of this NP is *inhibition*, but it bears no relevance – *telles* agrees with a more local noun *cytokines*, and that's the hypernym of the NPs that immediately follow *que*.

3. Surface recognition:

   FR: • *Maladies_fp de l'œ sophage et autres* **facteurs_mp** *qui retardent le transit_ms œ sophagien* **tels_mp que** *sténose et achalasie_fs.*

   EN: • *Abnormalities of the oesophagus and other* **factors** *which delay oesophageal emptying* **such as** *stricture or achalasia.*

   From a purely structural perspective, there are multiple possible ways of parsing the above sentence pair, different scopes with conjunction and subordination. But the reading with *facteurs* (mp) as hypernym is the only one that makes sense in natural language and is also the only one that *tels* in the FR sentence allows for given our hypothesis. With our heuristic, there is no need for any deep analysis beyond some shallow morphological tagging.

In the news commentary data, there are more instances with *tq* followed by one single hyponym. All of the 3 instances where our hypothesis failed to predict the correct hypernym in FR fall into this class – in one case, *tel-* seems to agree with the hyponym, in another, it doesn't agree with anything in the sentence (possible human errors). The third case, shown below, exemplifies the limit of this approach – if multiple nouns agree with *tq*, morphology alone cannot always uniquely detect the correct hypernym. It should be noted, however, that we only observed this pattern once in our sample of 100 sentences. Here, *guerre* "war" should be the hypernym, but *usure* "attrition" was predicted:

FR: *Malheureusement, les Israéliens_mp peuvent leur prouver qu'ils ne réussiront pas à détruire Israël sans faire l'expérience_fs d'une guerre_fs d'usure_fs violente* **telle_fs que** *celle qu'ils mènent.*

EN: *Unfortunately, the Israelis can show the Arabs that they cannot destroy Israel only by enduring the violent war of attrition that the Arabs are pursuing.*

There are also more cases where a pronominal NP (e.g. *celle* "the one") follows *que*, as in the example above as well as in ... *des expériences telles que celle de l'Iran* ...

[8] 0.5 due to partial alignment in 1 multiword expression.

	total number of unique sentences evaluated	hypothesis correct	FR hypernym extractable (using morfette)	EN hypernym extractable (using GIZA++)	
				best	worst
EMEA	50	50 (100%)	45 (90%)	41.5/44 (94.32%)	41.5/50 (83%)
WMT14 news commentary	50	47 (94%)	46 (92%)	46 (92%)	

Table 1: Results (*best* indicates when sentences misaligned to the extent that the relevant pattern is not available in EN are excluded, *worst* is when these cases are included in the count of total evaluated)

(lit.: "... experiences such as the one of Iran ..."), EN: ... *Iran's experiences* ..., but these are rarely translated into EN literally. In these sentences with only one hyponym, our hypothesis is still able to account for the majority of them. To obtain higher accuracy, one may want to restrict the hypothesized pattern to an enumeration of NPs following *que*, as opposed to "one or more NPs".

5. Conclusion

We presented a lexico-morphological pattern that facilitates accurate extraction of hypernyms in French, leveraging a morphological cue that enables unique disambiguation of many instances which would require semantic knowledge in other languages such as English. Transferring the results of this approach from French to English with the aid of bitexts and word alignment was found to be feasible. We understand that this discovery has its limit in coverage. That said, we hope that our effort in bringing this somewhat latent observation to light will inspire researchers to pay heed to these often neglected features in language that are available with little to no cost.

6. Bibliographical References

Bosma, W. and Vossen, P. (2010). Bootstrapping language neutral term extraction. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Chrupala, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with morfette. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Fu, R., Qin, B., and Liu, T. (2013). Exploiting multiple sources for open-domain hypernym discovery. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Seattle, Washington, USA, October. Association for Computational Linguistics.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lefever, E., Van de Kauter, M., and Hoste, V. (2014). Evaluation of automatic hypernym extraction from technical corpora in English and Dutch. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Ritter, A., Soderland, S., and Etzioni, O. (2009). What is this, anyway: Automatic hypernym discovery. In *Learning by Reading and Learning to Read, Papers from the 2009 AAAI Spring Symposium, Technical Report SS-09-07, Stanford, California, USA, March 23-25, 2009*, pages 88–93.

Shinzato, K. and Torisawa, K. (2004). Acquiring hyponymy relations from web documents. In Daniel Marcu Susan Dumais et al., editors, *HLT-NAACL 2004: Main Proceedings*, pages 73–80, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, et al., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia.

Tjong Kim Sang, E., Hofmann, K., and de Rijke, M. (2011). Extraction of hypernymy information from text. In *Interactive Multi-modal Question-Answering. Theory and Applications of Natural Language Processing*, pages 223–245. Springer, June.

Tjong Kim Sang, E. (2007). Extracting hypernym pairs from the web. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 165–168, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yamada, I., Torisawa, K., Kazama, J., Kuroda, K., Murata, M., De Saeger, S., Bond, F., and Sumida, A. (2009). Hypernym discovery based on distributional similarity and hierarchical structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 929–937, Stroudsburg, PA, USA. Association for Computational Linguistics.