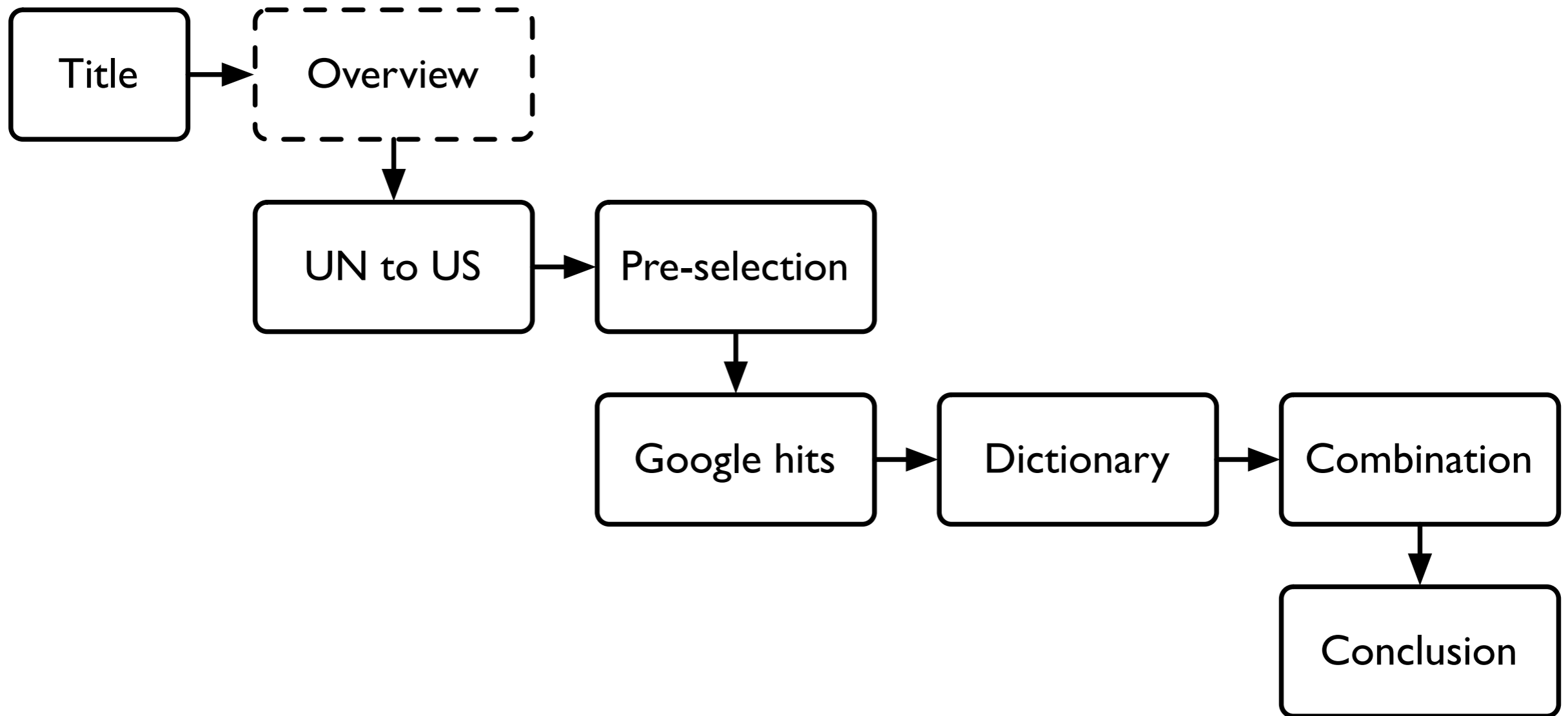


a time-saving approach to ontology mapping

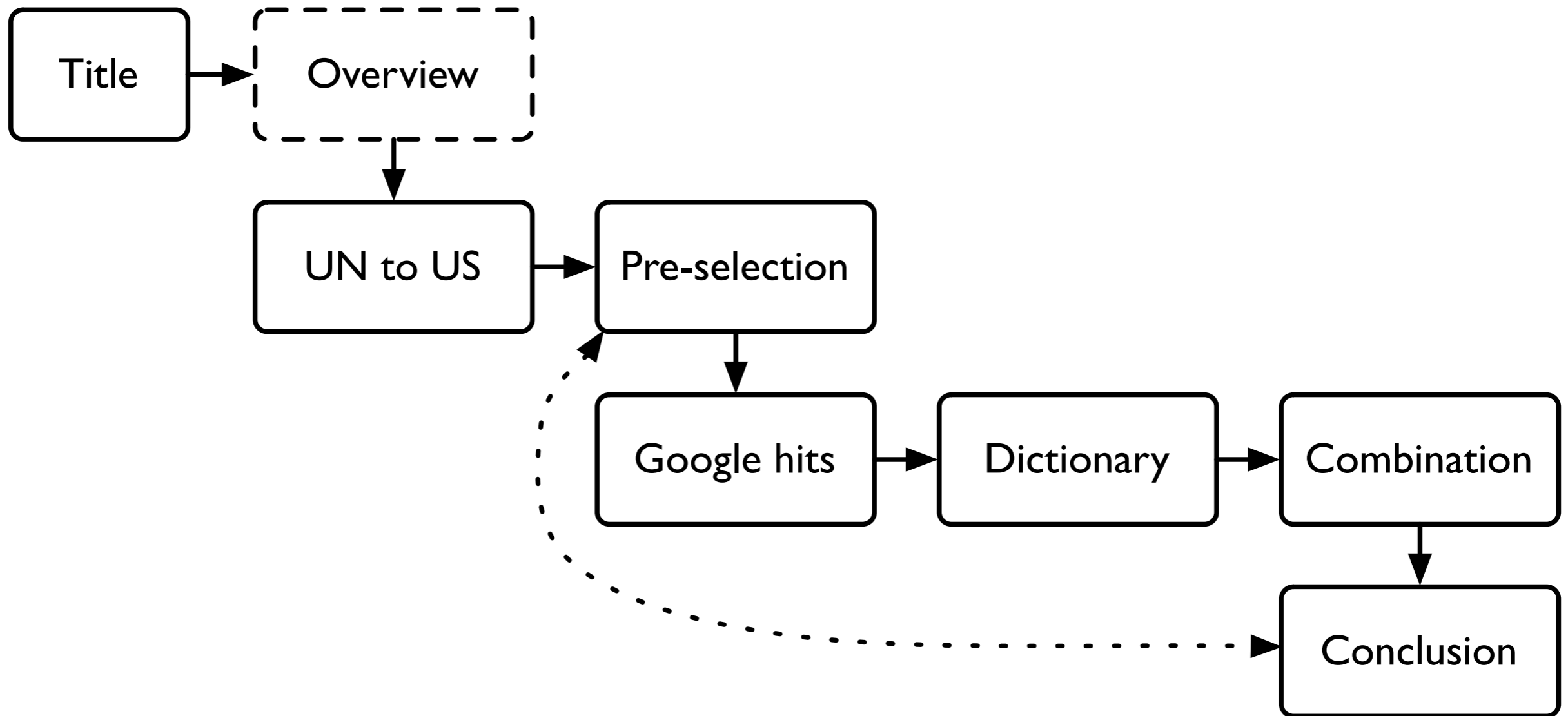
Willem Robert van Hage



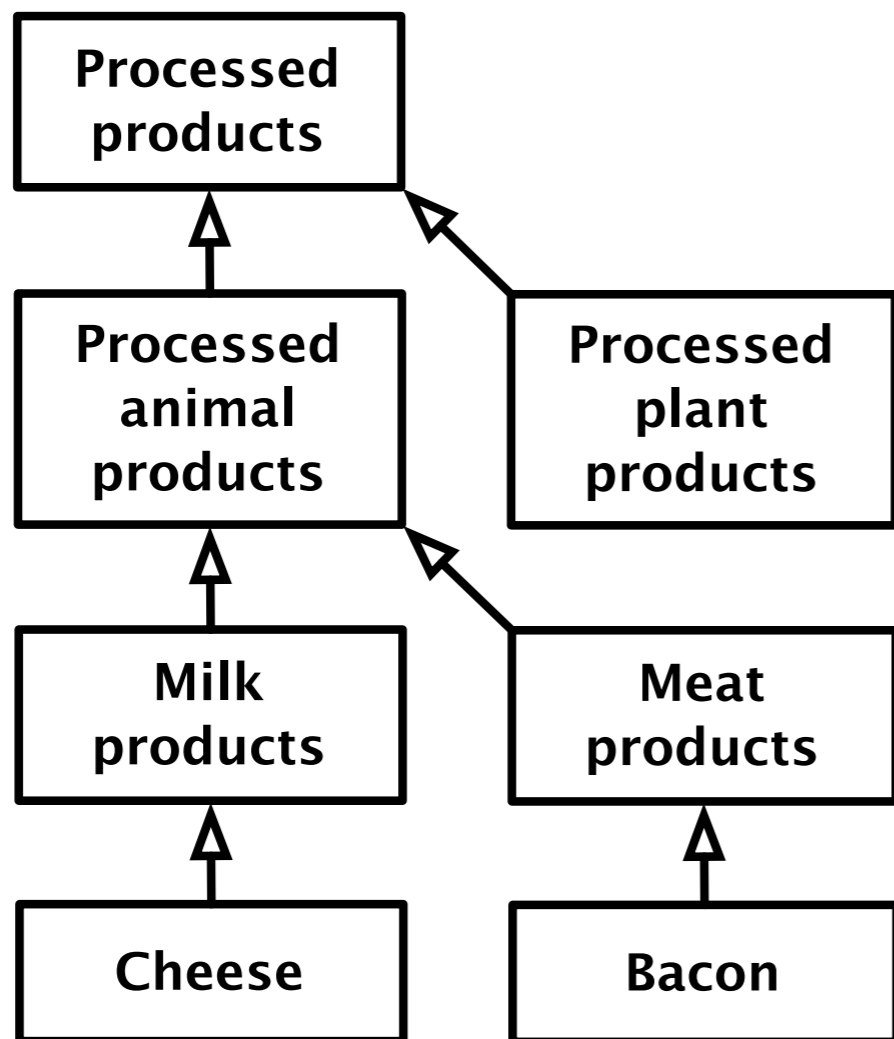
Overview



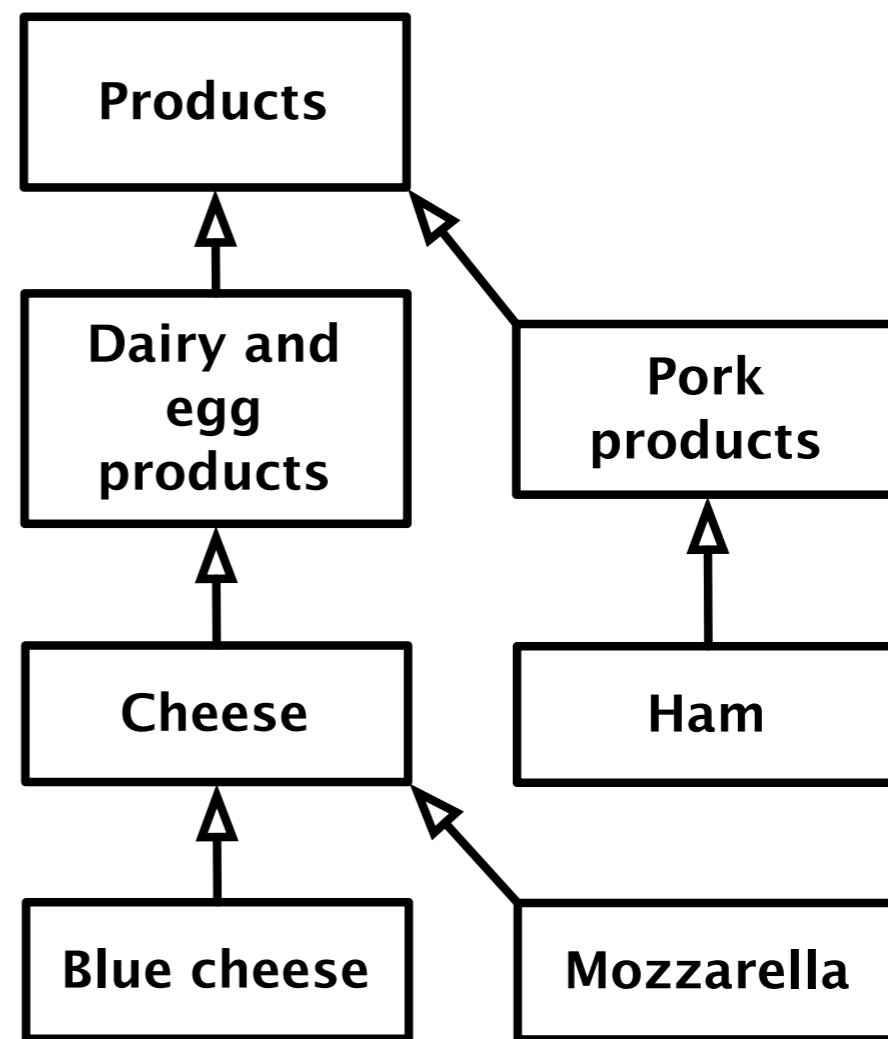
Overview



Mapping US to UN food vocabulary

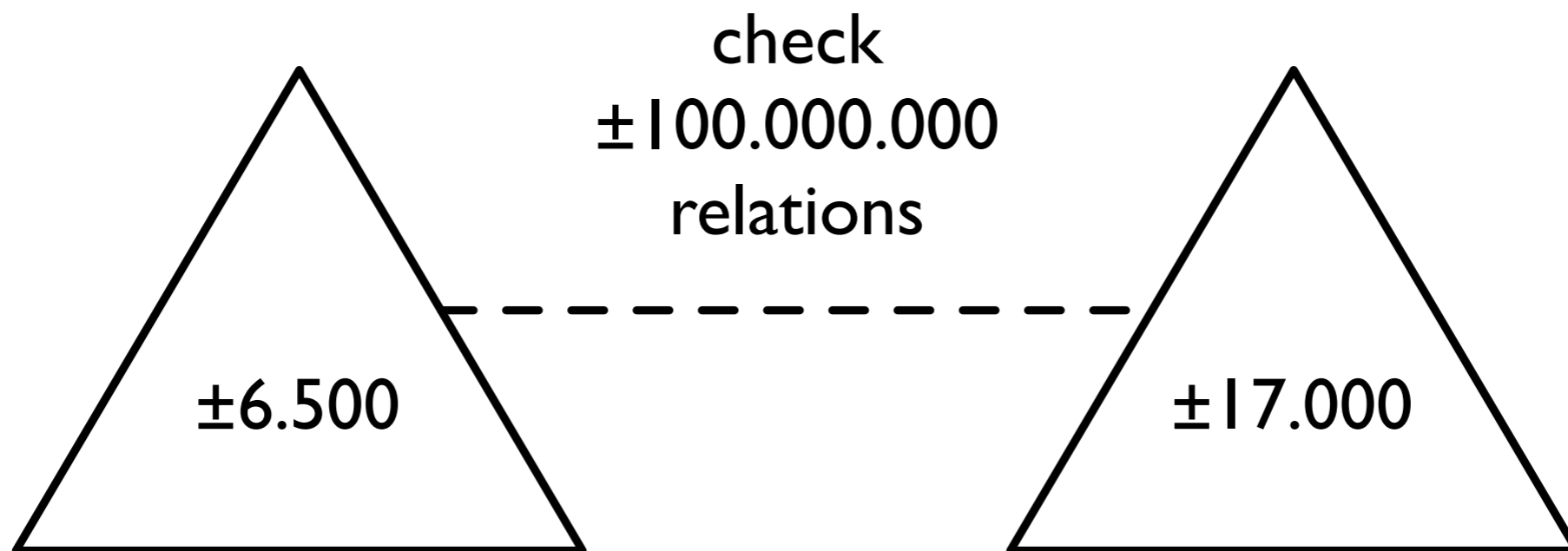


USDA SR-16 (US)

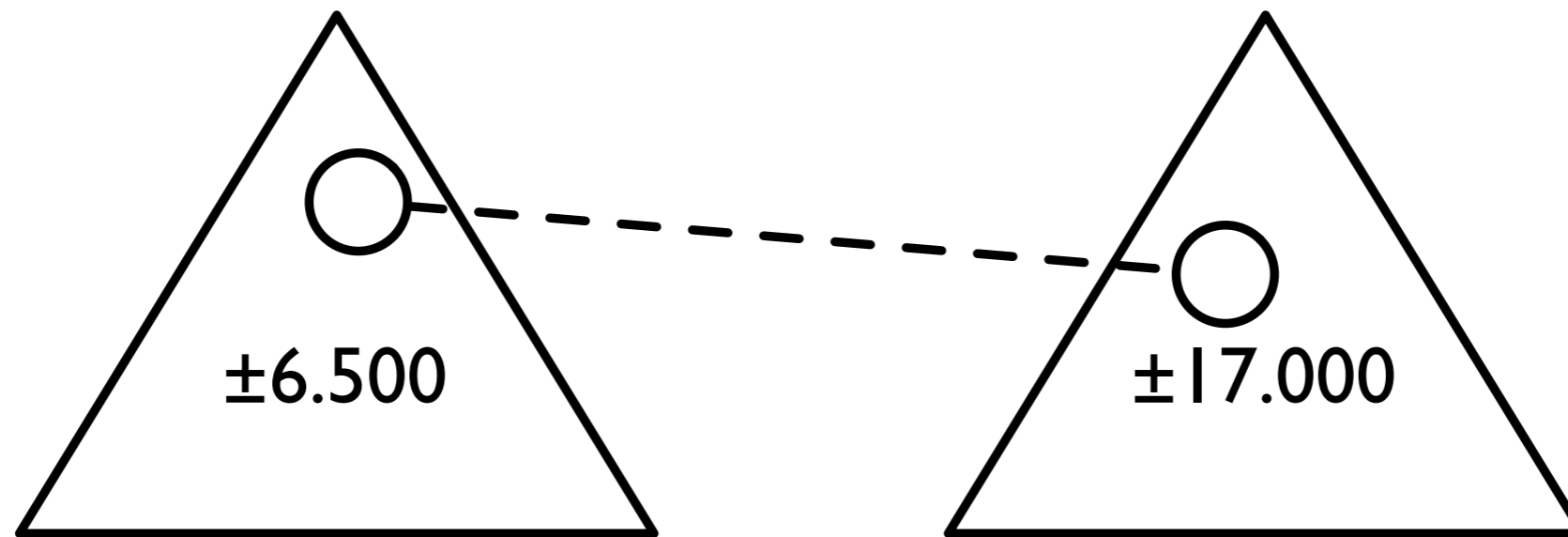


FAO AGROVOC (UN)

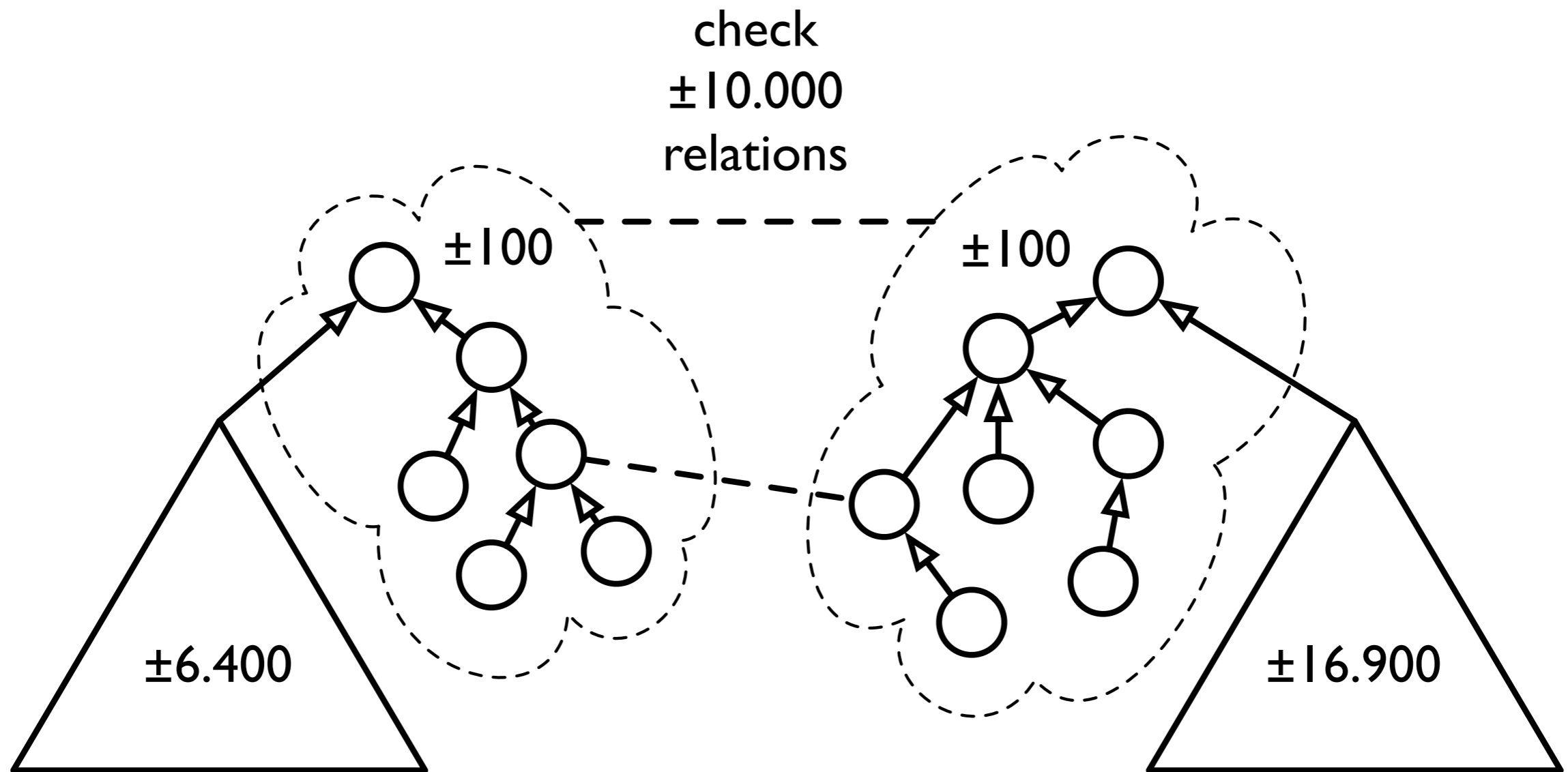
Pre-selection



Pre-selection



Pre-selection



Pre-selection

- Starting points must be of high quality (high Precision)
 - Not many are needed
 - False matches waste time
- Mapping of selections can be done extensively (high Recall)
 - Relatively little work
 - High correct match density

Pre-selection

- Starting points must be of high quality (high Precision)
 - Not many are needed
 - False matches waste time
- Mapping of selections can be done extensively (high Recall)
 - Relatively little work
 - High correct match density

Conclusion

- Pre-selection Precision around 90% guarantees good starting points with minimal manual correction
- Mapping these selections yields about 50% of all good mappings

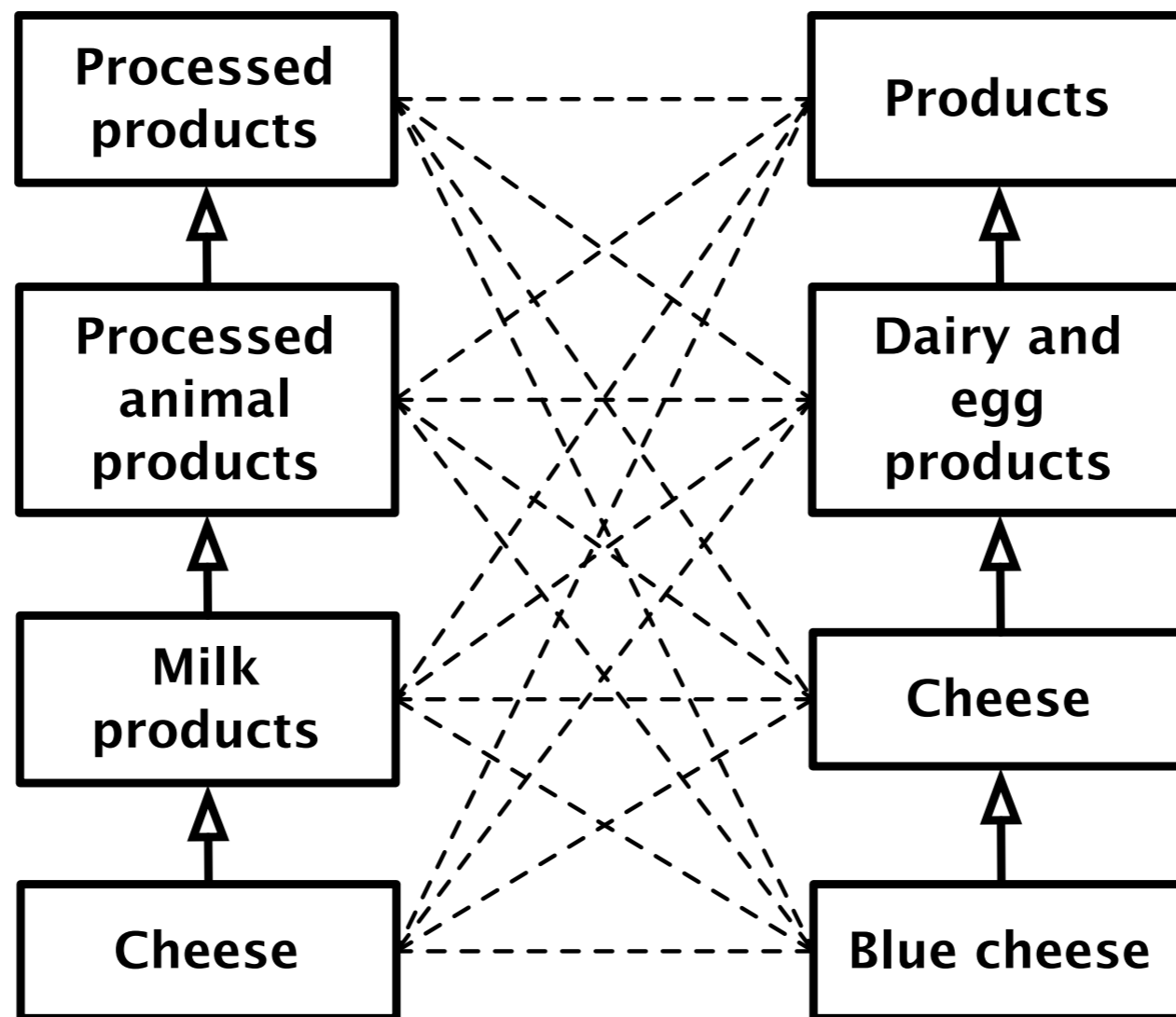
Pre-selection

- Starting points must be of high quality (high Precision)
 - Not many are needed
 - False matches waste time
- Mapping of selections can be done extensively (high Recall)
 - Relatively little work
 - High correct match density

Hearst patterns & Google hits

1. Select candidate pairs of concepts
2. Construct Google queries containing Hearst patterns for each pair
3. Send the queries to Google
4. Collect the hit counts and interpret a higher count as more evidence for the existence of a relation
5. Accept all candidates with enough evidence

Hearst patterns & Google hits



Hearst patterns & Google hits

1. Select candidate pairs of concepts
- 2. Construct Google queries containing Hearst patterns for each pair**
3. Send the queries to Google
4. Collect the hit counts and interpret a higher count as more evidence for the existence of a relation
5. Accept all candidates with enough evidence

Hearst patterns & Google hits

	<i>concept₁</i>	such as	<i>concept₂</i>
such	<i>concept₁</i>	as	<i>concept₂</i>
	<i>concept₁</i>	including	<i>concept₂</i>
	<i>concept₁</i>	especially	<i>concept₂</i>
<hr/>			
	<i>concept₂</i>	and other	<i>concept₁</i>
	<i>concept₂</i>	or other	<i>concept₁</i>

Hearst patterns & Google hits

1. Select candidate pairs of concepts
2. Construct Google queries containing Hearst patterns for each pair
- 3. Send the queries to Google**
4. Collect the hit counts and interpret a higher count as more evidence for the existence of a relation
5. Accept all candidates with enough evidence

Hearst patterns & Google hits

1. Select candidate pairs of concepts
2. Construct Google queries containing Hearst patterns for each pair
3. Send the queries to Google
- 4. Collect the hit counts and interpret a higher count as more evidence for the existence of a relation**
5. Accept all candidates with enough evidence

Hearst patterns & Google hits

meat \supseteq beef	839
oil \supseteq olive oil	829
meat \supseteq lamb	409
starch \supseteq rice	176
fruit \supseteq watermelon	143
pasta \supseteq macaroni	134

Hearst patterns & Google hits

1. Select candidate pairs of concepts
2. Construct Google queries containing Hearst patterns for each pair
3. Send the queries to Google
4. Collect the hit counts and interpret a higher count as more evidence for the existence of a relation
5. **Accept all candidates with enough evidence**

Hearst patterns & Google hits

Precision

Recall

similar level
concepts

17%

32%

low to high
level concepts

30%

53%

Hearst patterns & Google hits

- Why is mapping between low level concepts and high level concepts easiest?
- People rarely mention **is-a** relations, only when disambiguating with an example
- More data lead to better performance

Extraction from a dictionary

1. Find regularities in the dictionary that correlate with subclass relations
 - In this case the first head of a definition is usually a superclass
2. Select all entries that describe a concept from either ontology
3. Parse the entries
4. Extract the first head of each entry
5. Construct a subclass relation for each found superclass

Extraction from a dictionary

“**Basmati** - An aged, aromatic long-grain rice grown in the Himalayan foothills ...”

“**Bouillon** - A clear, delicately seasoned soup ...”

Extraction from a dictionary

1. Find regularities in the dictionary that correlate with subclass relations
 - In this case the first head of a definition is usually a superclass
- 2. Select all entries that describe a concept from either ontology**
3. Parse the entries
4. Extract the first head of each entry
5. Construct a subclass relation for each found superclass

Extraction from a dictionary

1. Find regularities in the dictionary that correlate with subclass relations
 - In this case the first head of a definition is usually a superclass
2. Select all entries that describe a concept from either ontology
- 3. Parse the entries**
4. Extract the first head of each entry
5. Construct a subclass relation for each found superclass

Extraction from a dictionary

1. Find regularities in the dictionary that correlate with subclass relations
 - In this case the first head of a definition is usually a superclass
2. Select all entries that describe a concept from either ontology
3. Parse the entries
4. **Extract the first head of each entry**
5. Construct a subclass relation for each found superclass

Extraction from a dictionary

1. Find regularities in the dictionary that correlate with subclass relations
 - In this case the first head of a definition is usually a superclass
2. Select all entries that describe a concept from either ontology
3. Parse the entries
4. Extract the first head of each entry
5. **Construct a subclass relation for each found superclass**

Extraction from a dictionary

salad ⇨ aemono

seaweed ⇨ agar agar

tooth ⇨ al dente

drink ⇨ ale

fish ⇨ anchovy

rice ⇨ basmati

Extraction from a dictionary

Precision

relations in general

53%

relations between
AGROVOC and SR-16

75%

Extraction and Google hits combined

1. Extract relations as before
2. Treat each extracted relation as a candidate relation in the Google hits method
3. Discard all relations that get insufficient evidence using the Google hits method

Extraction and Google hits combined

1. Extract relations as before
2. Treat each extracted relation as a candidate relation in the Google hits method
3. Discard all relations that get insufficient evidence using the Google hits method

Extraction and Google hits combined

meat \supseteq beef	839
oil \supseteq olive oil	829
...	...
rice \supseteq basmati	185
...	...
fruit \supseteq pepper	1
cold \supseteq mayonnaise	1

Extraction and Google hits combined

1. Extract relations as before
2. Treat each extracted relation as a candidate relation in the Google hits method
3. Discard all relations that get insufficient evidence using the Google hits method

Extraction and Google hits combined

Precision

relations in general

84%

relations between
AGROVOC and SR-16

94%

Conclusion

- Pre-selection Precision around 90% guarantees good starting points with minimal manual correction
- Mapping these selections yields about 50% of all good mappings