Comparison Mining

Youngjoong Ko
IS Laboratory, Dong-A University

New York trip

How about a Comparative Sentence like that?

Dahyun liked the Statue of Liberty more than the Empire State Building.

Dahyun loved the statue of liberty. Dahyun liked the Empire State building.

Introduction (1/2)

Comparison

- One of the most convincing ways of evaluation
- In many areas, comparisons have a great influence on decision making.

- These days, many web search engines are helping people look for their interesting entities.

However, directly reading each pages or reviews from the Web might not be a perfect solution.

- Small number of documents → a biased point of view
- Large number of documents → a time-consuming job

Introduction (2/2)

A comparison mining system

- automatically provides a summary of comparisons between two (or more) entities from a large amount of web documents
- very useful in many areas such as marketing

Our goal

- To build a Korean comparison mining system

Overview

iPhone vs. Galaxy-S

Which one is better?

Property	Comparative Type	%
Sound quality	Equality	?? %
	iPone is better.	?? %
	Galaxy-S is better.	?? %
Design	Similarity	?? %
	…..	?? %

System Developing Flow (4 stages)

1. Extracting comparative sentences from Texts → **2. Classifying comparative sentences into different types** → **3. Mining comparative Entities and Predicates** → **4. Analysis & Summary**

1. Extracting comparative sentences from texts

Problem Definition

Our basic idea is a keyword search.

- In many cases, comparative sentences contain some clue words, e.g., 보다 ([bo-da]: than)".
- A Korean linguist, *Ha* (1999), collected dozens of Korean comparative keywords with a linguistic perspective.
- Referring his research, we easily build an linguistic-based keyword set as follows:

  K_{ling} = { "같 ([gat]:same)", "보다 ([bo-da]: than)", "가장 ([ga-jang]: most)", … }

However, any method that depends on just these linguistic-based keywords has obvious limitations as follows:

- *K_{ling} is insufficient to cover all the actual comparison expressions*
- *There are many non-comparatives that contain some keywords.*

Our strategy

Our strategy is to first expand comparative keyword set and detect *Comparative Sentence candidates* (CS-candidates) by using them, and then eliminate non-comparative sentences from these candidates.

Step1.Detecting CS-candidates by keywords from texts

- At this time, recall↑ but precision↓

Step2. Eliminating non-comparatives from these candidates

- Finally, recall↑ and precision↑

Comparison lexicon (1/4)

To detect CS-candidates, we build up a comparison lexicon as follows:

Comparison Lexicon = K_{ling} U {Additional keywords that are frequently used for actual comparative expressions}

The lexicon consists of three parts as follows:

1. The elements of K_{ling} and their synonyms
   - e.g., “같 ([gat]: same),” “똑같 ([ddok-gat]: same),” “동일하 ([dong-il-ha]: same),” , “비슷하 ([bi-seut-ha]: similar),” “엇비슷하 ([eot-bi-seut-ha]: similar),” “유사하 ([yu-sa-ha]: similar),” ...

Comparison lexicon (2/4)

2. Idioms
   - e.g., “X의 손을 들어 주~ [X-eui son-eul deul-eo-ju~]” literally means “raise the hand of X” while it actually means “the winner is X.”
   - “시장 점유율 경쟁에서 신은 삼성의 손을 들어주었다. ([si-jang jeom-yu-yul kyeong-jang-e-seo sin-eun sam-sung-eui son-eul deul-eo-ju-eoss-da]): In the market share competition, God raised the hand of Samsung.”

     → This sentence can be interpreted as “*Samsung accounted for the largest share of the market*.”

Comparison lexicon (3/4)

3. Long-distance-words sequences
   - e.g., “<X는 [X-neun], 지만 [ji-man], Y는 [Y-neun], 다 [da]>”
     - This sequence means that the sentence is formed as < S(X) + V + but + S(Y) + V > in English (S: subject phrase; V: verb phrase; X, Y: proper nouns).
   - 쇼핑몰 X는 수수료 없이 전액 환불을 보장하지만, 쇼핑몰 Y는 환불 수수료를 요구한다. ([syo-ping-mol-X-neun su-su-ryo eops-i jeon-ak hwan-bul-eul bo-jang-ha-ji-man, syo-ping-mol-Y-neun hwan-bul su-su-ryo-reul yo-gu-han-da]: Shopping Mall X guarantees no fee full refund, but Shopping Mall Y requires refund-fee)”

     → This sentence does not directly compare two shopping malls. It implicitly gives a hint that X is more beneficial to use than Y.
   - We could regard a word, “지만 ([ji-man]: but),” as a single keyword. However, this word captures too many non-comparative sentences; the precision value can drop too low.

     → By using long-distance-words sequences, we can keep the precision value from dropping seriously low.

Comparison lexicon (4/4)

The comparison lexicon finally has a total of 177 elements.

- It is a time-consuming job but one time effort.
- We call each element "CK."
- Note that, our lexicon does not include any comparative/superlative POS tags
  - Unlike English POS taggers, there is no Korean comparative/superlative tags from POS tagger commonly.

Our lexicon covers 95.96% of the comparative sentences in our corpus.

- It means that we can successfully detect CS-candidates using the lexicon.

Non-comparatives (which are included in the CS-candidates)

Although using the lexicon shows a high recall, it shows relatively low precision of 68.39%.

- Comparison lexicon captures many non-comparative sentences also.
- e.g., "내일은 주식이 오를 것 같다." ([nai-il-eun ju-sik-i o-reul-geot gat-da]: I think stock price will rise tomorrow.)
  - This sentence is a non-comparative sentence even though it contains a CK, "같[gat]". This CK generally means "*same*," but it often expresses "*conjecture*."

Eliminating non-comparatives

To classify CS-candidates into two classes, *Comparative or Non-comparative*, we employ machine learning techniques.

- We did experiments with Naïve Bayesian classifier, Maximum Entropy classifier, Support Vector Machines (SVM), etc. Among these classifiers, SVM showed the best performance. Thus, we chose SVM as the proposed learning technique.

As features, we use the sequences of "continuous POS tags sequences within a radius of 3 words from each CK."

- Each feature has the form of "X → y"
  - "X" means a sequence of POS tags and "y" means a class (y_1: comparative, y_2: non-comparative)
- As CKs play the most important role, they are represented as a combination of th eir lexicalization and POS tag, e.g., "같/pa."
  - Here, "pa" is a POS tag whose meaning is "the stem of an adjective"

Experiments (1/3)

- Our corpus by three trained annotators (Kappa value: 0.85)

Total Sentences	Comparative	Non-comparative
7,389	2,383 (32.3%)	5,001 (67.7%)

- Evaluation measure: precision, recall, f1-score
- We did 5-fold cross validation (for every stage).

Experiments (2/3)

- **Example Sentence; the labels such as “ncn”, “jcs” are Korean POS tags**

Sentence	“비가 올 것 같다.” ([bi-ga ol geot gat-da]: I think it will rain.)
Class	Non-comparative
CK	같[gat]
Lexicalization/POS tag	비/ncn 가/jcs 오/pv ㄹ/etm 것/nbn 같/pa 다/ef ./sf

- **Examples of various features which we experimented with (%)**

Feature	Feature examples	Precision	Recall	F1-score
Lexical Unigram	비, 가, 올, 것, 같, 다	87.86	72.57	79.49
Lexical bigram	비가, 가올, 올것, 것같, 같다	80.15	68.26	73.73
Lexical Sequence (radius 1)	<것 같/pa 다>, <것 같/pa>, <같/pa 다>	87.06	87.65	87.35
POS tags sequence (radius 1)	< nbn 같/pa ef>, <nbn 같/pa>, <같/pa ef>	*87.65*	*88.74*	*88.19*

→ We achieved the best performance when using POS tags sequences.

Experiments (3/3)

- **Comparison of five radius options (%)**

Radius option	Feature examples	Precision	Recall	F1-score
r1	<nbn 같/pa ef>, <nbn 같/pa>, <같/pa ef>	87.65	88.74	88.19
r2	<etm nbn 같/pa>, < nbn 같/pa ef>, <같/pa ef sf>, <nbn 같/pa>, <같/pa ef>	90.47	88.56	89.50
r3	**<pv etm nbn 같/pa ef sf>, <pv etm nbn 같/pa ef >, <etm nbn 같/pa ef sf>, …, <같/pa ef>**	***92.24***	***88.31***	***90.23*** ⇦
r4	<jsc pv etm nbn 같/pa ef sf>, <pv etm nbn 같/pa ef sf>, …, <같/pa ef>	93.48	87.09	90.17
r5	<ncn jsc pv etm nbn 같/pa ef sf>, <pv etm nbn 같/pa ef sf>, …, <같/pa ef>	94.24	86.35	90.10

→ Based on f1-score, r3 showed higher performance than r1 and r2; statistically significant at $p<0.01$. On the other hand, although the r3 was slightly higher than r4 and r5, there was not statistically significant difference. Hence, among r3, r4 and r5, we chose r3, the shortest one, as our final radius option.

Summary of comparative sentence extraction

We first detected CS-candidates with the comparison lexicon. Then we eliminated non-comparatives from the candidates using a machine learning technique (method: SVM, feature: continuous POS tags sequences within a radius of 3 words from each CK).

As a result, we could extract comparative sentences from texts with the significant performance, an f1-score of 90.23%.

2. Classifying comparative sentences into different types

Type Definition

We define seven types for comparative sentences;

Type	Sentence Example	View
1) Equality	"X와 Y는 디자인이 같다." (X and Y have the same design.)	From a Linguistic View
2) Similarity	"X의 디자인은 Y하고 비슷하다." (The design of X is similar to that of Y.)	
3) Difference	"X는 그 점에서 Y와 다르다." (X differs from Y on that point.)	
4) Greater or lesser	"X차는 승차감이 Y차보다 부드럽다." (Car X gives a smoother ride than Car Y.)	
5) Superlative	"후보들 중에서 X가 가장 신뢰가 간다." (X is the most reliable among the candidates.)	
6) Pseudo	"폰 X는 실용적이라기보다 장식용이다." (Phone X is ornamental rather than practical.)	
7) Implicit	"쇼핑몰 X는 수수료 없이 전액 환불을 보장하지만, 쇼핑몰 Y는 환불 수수료를 요구한다. (Shopping Mall X guarantees no fee full refund, but Shopping Mall Y requires refund-fee.)	With an enhanced view

Problem Definition (1/2)

We decide to first use the comparison lexicon to do this task as we used it in the previous stage.

- **We easily match each CK to a particular type, e.g., "같 ([gat]: same)" to "*1) Equality*", "보다 ([bo-da]: than)" to "*4) Greater or lesser*".**
- **Then, we simply match each sentence to a particular type based on the CK types; e.g., a sentence which contains the word "가장 ([ga-jang]: most)" is matched to "*5) Superlative*" type.**

However, classifying comparative sentences using just the CK information has a serious limitation as follows:

- ***There is no one-to-one relationship between keyword types and sentence types.***

Problem Definition (2/2)

For example, although we easily match the CK "보다 ([bo-da]: than)" to "*Greater or lesser*" without doubt, we observe that the type of CK itself does not guarantee the correct type of the sentence as we can see in the following three sentences:

1) **"X의 품질은 Y<u>보다</u> 좋지도 나쁘지도 않다." ([X-eui pum-jil-eun Y-<u>bo-da</u> jo-chi-do na-ppeu-ji-do an-ta]: The quality of X is neither better nor worse <u>than</u> that of Y.)**
   → It can be interpreted as "The quality of X is similar to that of Y." (*Similarity*)
2) **"X가 Y<u>보다</u> 품질이 좋다." ([X-ga Y-<u>bo-da</u> pum-jil-I jo-ta]: The quality of X is better <u>than</u> that of Y.)**
   → It is consistent with the CK type (*Greater or lesser)*
3) **"X는 다른 어떤 카메라<u>보다</u> 품질이 좋다." ([X-neun da-reun eo-tteon ka-me-ra-bo-da pum-jil-i jo-ta]: X is better <u>than</u> any other cameras in quality.)**
   → It can be interpreted as "X is the best camera in quality." (*Superlative*)

Error-driven approach

Like the previous comparative sentence extraction task, we also conduct experiments for type classification using the same features (continuous POS tags sequences within a radius of 3 words from each CK) and the same learning technique (SVM).

- **Here, we achieved an accuracy of 73.64%.**

Next, we tested a completely different technique, the Transformation-Based Learning (TBL) method.

- **TBL is well-known to be relatively strong in sparse problems.**
- **We observed that the performance of type classification can be influenced by very subtle differences in many cases.**
- **Hence, we think that an error-driven approach can perform well in comparative type classification.**
- **Experimental results showed that TBL actually performed better.**

Transformation Rules (1/2)

The transformation rules are generated on the basis of the following templates.

Change the type of the current sentence from ***x*** to ***y*** if this sentence holds the CK of ***k***, and …
1. the preceding word of ***k*** is tagged ***z***. 2. the following word of ***k*** is tagged ***z***. 3. the second preceding word of ***k*** is tagged ***z***. 4. the second following word of ***k*** is tagged ***z***. 5. the preceding word of ***k*** is tagged ***z***, and the following word of ***k*** is tagged ***w***. 6. the preceding word of ***k*** is tagged ***z***, and the second preceding word of ***k*** is tagged ***w***. 7. the following word of ***k*** is tagged ***z***, and the second following word of ***k*** is tagged ***w***.

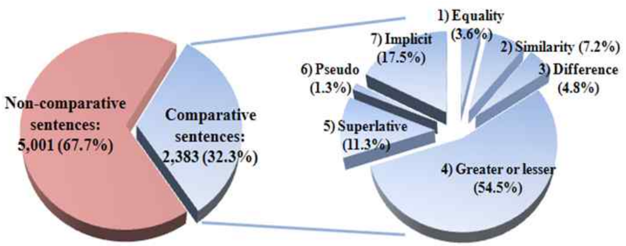
Transformation Rules (2/2)

Rule example

- **e.g.**, "Change the type of the current sentence from "***Greater or lesser***" to "***Superlative***" if this sentence holds the CK of "보다 ([bo-da]: than)", and the second preceding word of the CK is tagged as ***mm***" is a transformation rule generated by the third template.

Experiments (1/2)

- Distribution of the corpus



- Evaluation measure: Accuracy

Experiments (2/2)

- Evaluation of threshold option (%); threshold n means that the learning iterations continues while C_i-$E_i \geq n+1$

Threshold	Accuracy (%)
Threshold = 0 (C_i - $E_i \geq 1$)	79.99
Threshold = 1 (C_i - $E_i \geq 2$)	**81.67** ⇦
Threshold = 2 (C_i - $E_i \geq 3$)	80.34

→ We achieved the best result when we used a threshold value of 1. The accuracy difference was statistically significant at $p < 0.01$.

Integrated results

We individually perform comparative sentence extraction (Stage 1) and type classification (Stage 2). To evaluate the effectiveness of our two-step processing, we performed one-step classification experiments using SVM and TBL; here, non-comparative sentences are regarded as the eighth type.

Processing		Accuracy (%)
One-step processing (classifying eight types at a time)	comparison lexicon & SVM	75.64
	comparison lexicon & TBL	72.49
Two-step processing (proposed)		***88.59***

Summary of comparative sentence classification

To effectively classify comparative sentences into seven types, we employed the TBL method.

If we only rely on the CK type, we could not solve the ambiguity problem of CK type. After investigating a large number of actual comparative sentences, we could find numerous ambiguous cases.

By using TBL method, we were able to improve performance; we could classify comparative sentences into seven types with an accuracy of 81.67%.

3. Mining comparative entities and predicates

Our goal

We extract comparative entities and predicates taking into account the characteristics of each type.

For example, from the sentence "Stock-X is worth more than stock-Y." belonging to "*Greater or lesser*" type, we extract "*stock-X*" as a subject entity (SE), "*stock-Y*" as an object entity (OE), and "*worth*" as a comparative predicate (PR).

In summary, our goal is to extract three kinds of comparative elements (two comparative entities (SE and OE) and one comparative predicate (PR)) from each comparative sentence.

- We only present the results of two types: "*Greater or lesser*" and "*Superlative*"; these two types cover 65.8% of whole comparative sentences. We will report the experimental results on the other five types soon.

Our strategy

Our strategy is to first detect *Comparative Element candidates* (CE-candidates), and then choose the answer among the candidates.

Step1.Detecting CE-candidates in each comparative sentence

Step2. Finding the answer among these candidates

Problem Definition (1/3)

At first, we decide to select each of noun words as a candidate for SE/OE, and each of adjective (or verb) words as a candidate for PR. However, this candidate detection has serious problems as follows:

- There are many actual SEs, OEs, and PRs that consist of multiple words.
- There are many sentences with no OE, especially among superlative sentences. It means that the ellipsis is frequently occurred in superlative sentences.

Problem Definition (2/3)

Multi-words element:

"X파이가 Y파이보다 싸고 맛있다. ([X-pa-i-ga Y-pa-i-bo-da ssa-go mas-it-da]: Pie X is cheaper and more delicious than Pie Y.)"

- "X파이 (Pie X)" is a SE, "Y파이 (Pie Y)" is an OE, and "싸고 맛있다 (cheaper and more delicious)" is a PR.

"대선 후보들 중 Z가 가장 믿음직하다. ([dai-seon hu-bo-deul jung Z-ga ga-jang mit-eum-jik-ha-da]: "Z is most trustworthy among the presidential candidates.)"

- "Z" is a SE, "대선 후보들 (the presidential candidates)" is an OE, and "믿음직하다 (trustworthy)" is a PR.

→ "싸고 맛있다 (cheaper and more delicious)" and "대선 후보들 (the presidential candidates)" are composed of multiple words.

Problem Definition (3/3)

Omitted OE:

"뭐니뭐니해도 Z폰이 제일 좋아! ([mwo-ni-mwo-ni-hae-do Z-pon-i choi-go-ya]: No matter what they say, Phone Z is best!)"

- "Z폰 (Phone Z)" is a SE, OE is omitted, and "최고 (best)" is a PR.

→ In our corpus, about 70% of the *Superlative* sentences don't have their OE.

Detecting CE-candidates

Preprocessing step for easy detection of CE-candidates:

Through this preprocessing step, we represent potential SEs/OEs as one "N" and potential PRs as one "P"

The following process is one of the simplification processes for making "N"

- **Change each noun (or each compound noun) into a symbol "N"**
- **Change "N + jxm (a postposition whose meaning is "of")+N" into one "N"**
- **…**

And, the following process is one of the simplification processes for making "P".

- **Change "pa (adjective)" and "pv (verb)" into a symbol "P"**
- **Change "P + ecc (a suffix whose meaning is "and") + P" into one "P"**
- **….**

→e.g., "cheaper and more delicious" is tagged as one "P" by the above two rules.

Finding the answer elements (1/2)

To find the answer among the detected CE-candidates, we employ machine learning techniques;

- **In our experiments, although both MEM and SVM showed outstanding performance, there was no significant difference. Hence, we only report the results of SVM.**

As features, we use the patterns that consist of POS tags, CKs, and "P"/"N" sequences within a radius of 4 POS tags from each "N" or "P".

Finding the answer elements (2/2)

- **Feature examples**

Example sentence	"X파이가 Y파이보다 싸고 맛있다." (Pie X is cheaper and more delicious than Pie Y.)
After POS tagging	X파이/nq + 가/jcs + Y파이/nq + 보다/jca + 싸/pa + 고/ecc + 맛있/pa + 다/ef +./sf
After preprocessing	X파이/N(SE) + 가/jcs + Y파이/N(OE) + 보다/jca + 싸고맛있다/P(PR) + ./sf
Features for SE (Pie X)	<N(SE), jcs, N, 보다/jca,P>, …, <N(SE), jcs>
Features for OE (Pie Y)	<N, jcs, N(OE), 보다/jca,P, sf>, …, <N(OE), 보다/jca >
Features for PR (cheaper and more delicious)	<N, jcs, N, 보다/jca,P(PR), sf>, …, <P(PR), sf>

Experiments (1/4)

- **Corpus: We use 460 comparative sentences (*Greater or lesser*: 300, *Superlative*: 160)**
- **Evaluation measure: Accuracy**
- **The portion of multiple-word comparative elements (%); we calculate without the Superlative sentences that do not have any OE.**

Element	Greater or lesser	Superlative
SE	30.0	24.4
OE	31.3	32.6
PR	8.3	8.1

→ As given above, each multiple-word portion, especially in SEs and OEs, is quite high. This fact proves that it is absolutely necessary to allow multiple-word comparative elements.

Experiments (2/4)

- Success rate in CE-candidate detection (%)

Element	Greater or lesser	Superlative
SE	95.3 (65.3)	98.1 (73.7)
OE	92.0 (60.7)	93.7 (24.4)
PR	98.3 (90.0)	98.7 (90.6)

→ The significant differences between before and after indicate that we successfully detect CE-candidates through the multi-words preprocesses.

Experiments (3/4)

- Comparison of five radius options (SE/OE/PR, %)

Radius option	Greater or lesser	Superlative
r1	82.61	76.98
r2	85.78	79.75
r3	88.32	80.47
r4	**89.45**	**81.88**
r5	89.20	81.76

→ r4 showed higher performance than r1, r2, and r3. r4 was also slightly higher than r5. We thus chose r4 as our final radius option.

Experiments (4/4)

- Final Results (%)

Element	Greater or lesser	Superlative	Total
SE	86.00	84.38	85.43
OE	89.67	71.25	83.26
PR	92.67	90.00	91.74
SE/OE/PR	89.45	81.88	***86.81***

Summary of comparative element mining

We first did multi-word merge preprocess and detected CE-candidates successfully. Then we chose the answer elements using a machine learning technique (method: SVM, feature: the patterns that consist of POS tags, CKs, and "P"/"N" sequences within a radius of 4 POS tags from each "N" or "P").

As a result, we could extract comparative entities and predicates from two types of comparative sentences ("*Greater or lesser*" and "*Superlative*") with an overall accuracy of 86.81%; we are studying the other five types and will report the results soon.

4. Analysis/Summary & References/Other Topics

This task is currently on the drawing board

The important point to note at this stage is that, although the comparative elements are different, they can have the same meaning as follows:

(1) "Phone X is lighter than Phone Y"

(2) "Phone Y is heavier than Phone X"

These two sentences have the same meaning but the extracted elements are different.

Element	(1)	(2)
SE	Phone X	Phone Y
OE	Phone Y	Phone X
PR	lighter	heavier

References and Other Topics

References

- Seon Yang and Youngjoong Ko, "Finding Relevant Features for Korean Comparative Sentence Extraction," Pattern Recognition Letters, Elsevier Science, Vol. 32, No. 2, pp. 293-296, January 2011.
- Seon Yang and Youngjoong Ko, "Extracting Comparative Entities and Predicates from Texts Using Comparative Type Classification, " ACL-HLT 2011, in Portland Oregon, USA, pp. 1636-1644, June, 19-24, 2011.
- Seon Yang and Youngjoong Ko, "Extracting Comparative Sentences from Korean Text Documents Using Comparative Lexical Patterns and Machine Learning Techniques, " ACL-IJCNLP 2009), in Singapore, pp. 153-156, Aug. 2-7, 2009.

Other Research Topics

- Opinion Mining: Sentiment Classification
- Text Mining: Text Classification/Summarization
- Information Retrieval : Equations retrieval
- NLP Oriented topics: Dialogue System (Speech-act analysis, Dialogue Modeling)