

SinK DaT (Sino-Korean Detector and Translator)

Reducing the Generation Gap

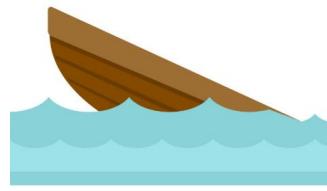
Team 9

20160697 Mina Huh, 20160793 SeokJun Kim,

20160811 Jeongeon Park, 20160832 Juhoon Lee, 20170410 Hyunchang Oh



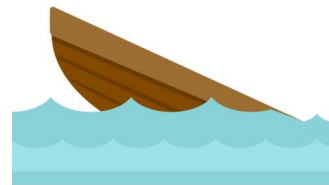
'Sino-Ko' Words



- Chinese/Hanja based words that are incorporated into the Korean language with Korean Pronunciation.

Hanja Character	Pronunciation	Meaning
學 +	'Hak'	배울 학, to learn
生 =	'Saeng'	날, to be born and living
學生 Combined	'Hak Saeng'	학생, student

'Sino-Ko' Words

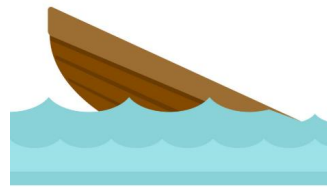


- There also exist Korean **homophones** that have different Chinese characters, hence different meanings

Hanja Character	Pronunciation	Meaning
醫師	'Eui- Sa'	Doctor
義士	'Eui- Sa'	Patriot

~70% of Korean words are 'Sino- Ko'

Sino-Ko word Illiterate Generation..

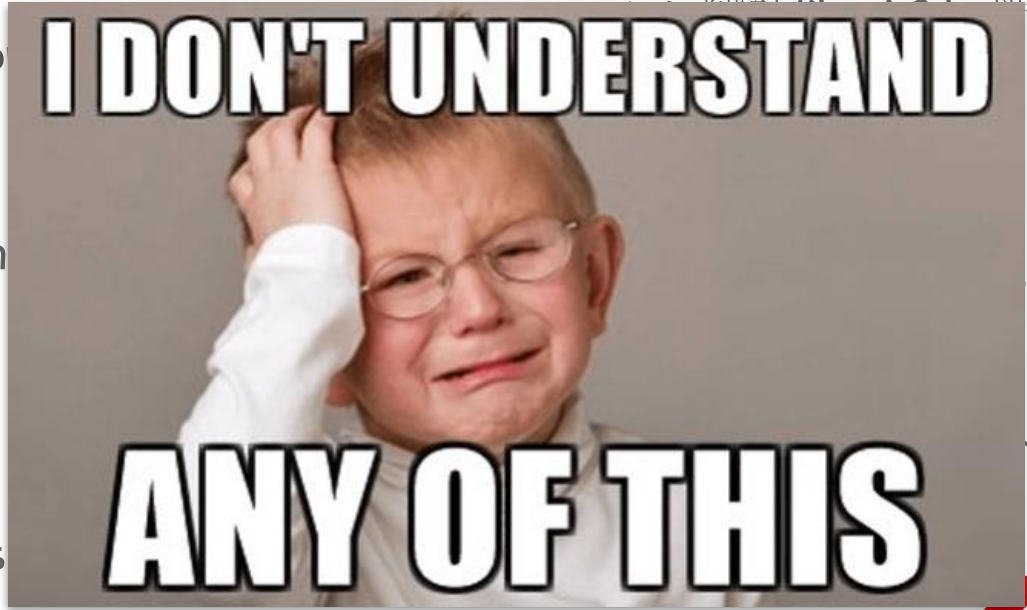


... (S) '陣痛, 鎮痛('Jin Tong)'

young generatio

I DON'T UNDERSTAND

71% of elem



ch text..

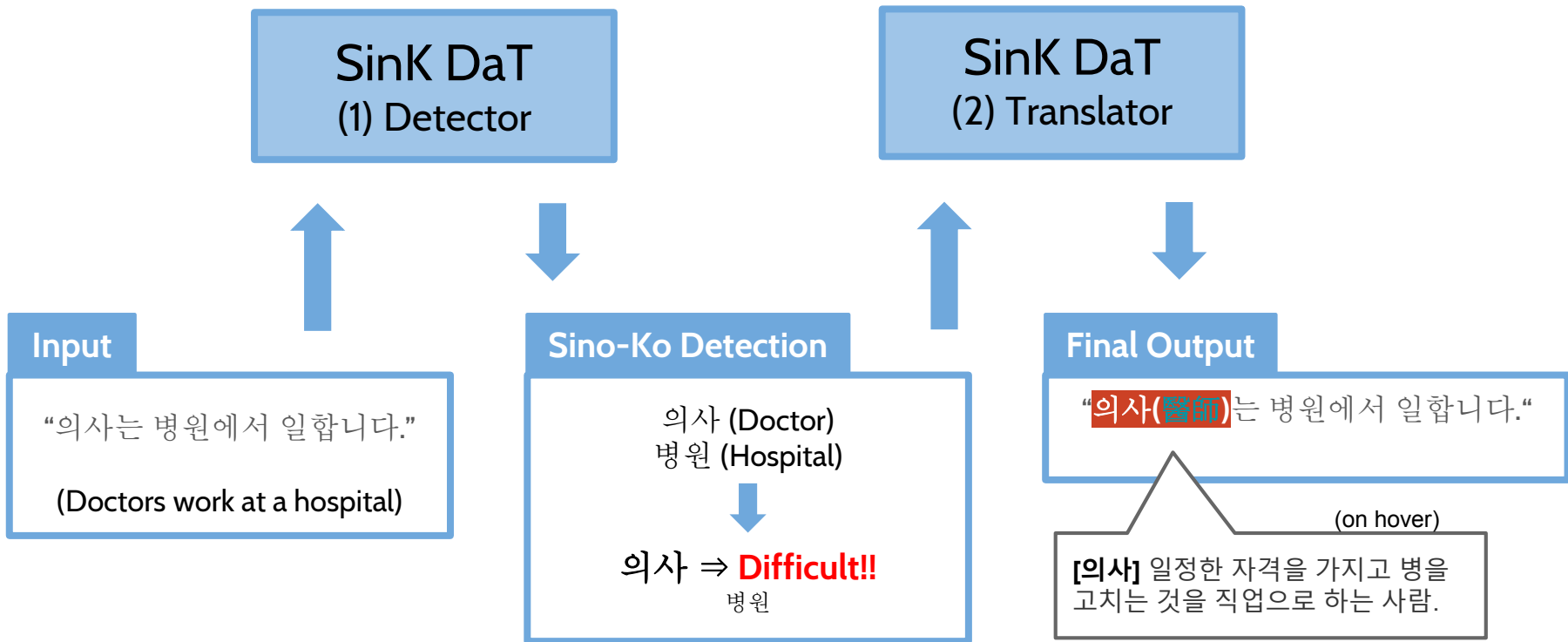
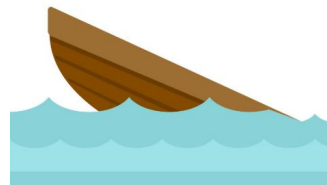
n Jung Geun?

Anonymous

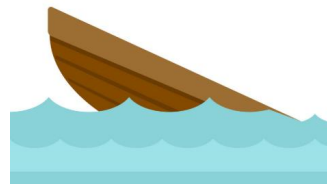


ChosunMedia
조선일보

SinK DaT : Context-based Sino-Ko Word Detector and Translator

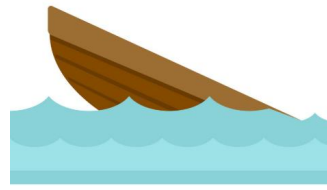


Technical Approach



- Dataset & Tools
- Preprocessing
- 'Sino-Ko' Word Detector
 - Identifying 'Sino-Ko' words
 - Scoring the difficulty of each word
- 'Sino-Ko' Word Translator
 - Replacing / defining the word: Word sense disambiguation

Dataset



Chosun Ilbo News Library 1990s news articles



National Institute of Korean Language

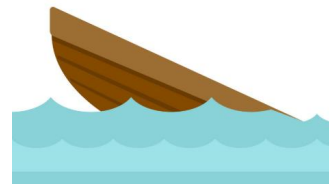


Kyohaksa - Korean Textbook Publisher



Hangum - Official Hanja Level Test

Tools



Language tools

KoNLTK

Sentence
Segmentation,
Word Tokenization
and POS tagging



Find
Word Definition
and Synsets

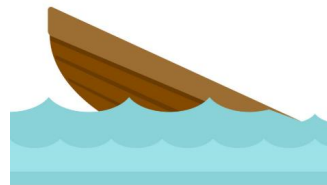
Dictionaries

N 국어사전

사전
한자

Find
Chinese Character
Definition and Level

Scoring the Difficulty of 'Sino-Ko' Words

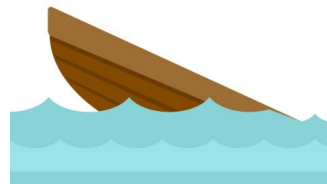


1. Count the usage of different Synsets

2. Analyze level of each Hanja 번뇌 (煩 + 惱)
level3 level3

3. Count the appearance in EASY / DIFFICULT
Corpus





Scoring the Difficulty of 'Sino-Ko' Words

1. Count the usage of different Synsets

번뇌1 (n | 14442671)

KorLex
↳ 정신1, 정신적 특징1
↳ 감정1, 기분2
↳ S 계박1, 번뇌1

유리1 (a | 02787734)

유리4 (n | 04865892)

유리2 (n | 10621413)

유리3 (n | 14033732)

유리5 (n | 14440794)

유리6 (n | 15301385)

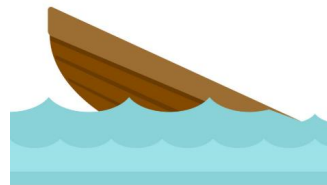
KorLex
↳ 행동1, 행위1
↳ 행동1
↳ 변경1, 변환1
↳ 완전성 변경1
↳ 분리1
↳ S 유리6

← DIFFICULT words tend to have less Synsets

EASY words tend to have many Synsets →



Scoring the Difficulty of 'Sino-Ko' Words



2. Analyze level of each Hanja 번뇌 (煩 + 惱) level3 level3

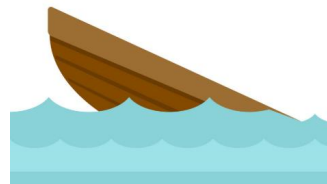
2020년도

한자급수자격검정시험 등급별 검정과목 및 출제형식

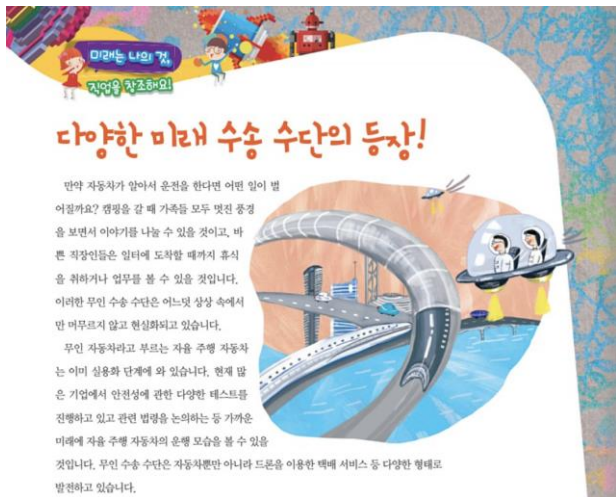
등급	검정과목	검정방법
대사범	· 대학 · 논어 · 맹자 · 중용 · 고문진보 · 사략 · 고급한문 II (선정단원) · 기타	필기시험
사범	· 한문지식(한자 5,000) · 고급한문(선정단원)	필기시험
1급	· 한문지식(한자 3,500) · 중급한문 II (선정단원)	필기시험
준1급	· 한문지식(한자 2,500) · 중급한문 I (선정단원)	필기시험
2급	· 한문지식(한자 2,000) · 초급한문 II (선정단원)	필기시험
준2급	· 한문지식(한자 1,500) · 초급한문 I (선정단원)	필기시험
3급	· 한문지식(한자 1,000)	필기시험



Scoring the Difficulty of 'Sino-Ko' Words



3. Count the appearance in EASY / DIFFICULT Corpus



Keumsung 6th grade Textbook



[단독] 원로들에게 속내 밝힌 文대통령 "최선 다했는데... 北에 광강히 실망"

- [단독] 철모 쓰고 총에 착검... 최전방 북한군 심상치않다
- '황' 순간 가루가 된 연락사무소... 北, 영상도 공개
- 北, 사흘전 폭탄 설치했다... 국방부 "건물서 불꽃 튀어"



김병기 "국정원, 대통령 보고 어떻게 했냐... 기망한거냐"

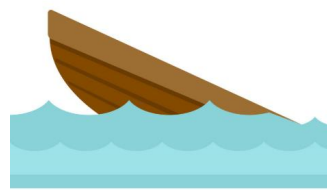
국회 정보위원회 여당 간사인 더불어민주당 김병기 의원은 17일 남북 관계 상황에 대해 국정보원 등 정보 당국이 청와대에 지나치게 낙관적인 보고를 했을 가능성을 언급했다. 국정원 출신인 김 의원은 이날 기자들과 만나 최근 남북공동연락사무소 폭파 등에 대해 "북..

대포로 안센게 어디냐" 송영길에... 통합당 사퇴 결의안 검토

· "송영길 발언은 부적절" 정의당마저 등돌리

Chosun Ilbo Headline

Word Sense Disambiguation



조선일보 건강

항생제 개발로 '죽음의 감염병' 물리쳤다

머크 매뉴얼 선정 '20세기의 10대 업적'

20세기 초 의사들의 왕진가방에는 비소와 수은 같은 독극물이 들어 있었다. 가장 많은 환자의 고통을 덜어주기 위해서다. 의사가 할 수 있는 일이라고 과사병(살이 썩는 병) 환자의 골-다리를 지른 뒤, 수술 자리가 나타나지 않기를 바라는 것 뿐이었다. 그러나 이제 의사들은 병든 심장이나 간을 바꿀지기 할 수 있을 정도가 됐다. 1899년 미국에서 초안이 발췌된 '머크 매뉴얼(Merck Manual)'은 전세계 의사들이 가장 많이 참고하는 교과서. 머크 매뉴얼의 10대 비소우 편집장은 '머크 매뉴얼' 21세기 개정판을 발간하며, 20세기 의학의 10대 업적을 정리했다.

1940년대 이후 이른바 '3대 항생제' (살모계열 항생제, 페니실린, 스트렙토마이신)가 환자의 치료에 사용되면서 각종 감염질환의 사망률이 크게 떨어졌다. 런던의대 출신의 미생물학자 알렉산더 플레밍은 1928년, 우연히 보드 상구균의 발생을 억제하는 진균(페니실)을 발견했으며, 1940년 영국에서 이를 안정된 분말로 정제하는데 성공했다. 이 약은 1944년 2차대전 말엔 군인을 치료하는데 최초로 사용되던 항

소아마비-천연두 등 백신 개발로 자취 감춰
효진단-치료 기술 발달... 초기 발견하면 원치도

1940년대 이후 이른바 '3대 항생제' (살모계열 항생제, 페니실린, 스트렙토마이신)가 환자의 치료에 사용되면서 각종 감염질환의 사망률이 크게 떨어졌다. 런던의대 출신의 미생물학자 알렉산더 플레밍은 1928년, 우연히 보드 상구균의 발생을 억제하는 진균(페니실)을 발견했으며, 1940년 영국에서 이를 안정된 분말로 정제하는데 성공했다. 이 약은 1944년 2차대전 말엔 군인을 치료하는데 최초로 사용되던 항

의사들은 암 환자에 대한 치료하는 기술이 가능해짐에 따라 신암 치료에 힘쓰고 있다.

“부자기관... 했다. 또... 재적으로... 재하는 외... 는 방안도... 인기업의... 하기 위해... 영하겠다”... 후닥터(고... 3200여 외... 인...)

1999년 11월 1일 월요일 42판 제2451호 25

11월 독립운동가 강우규義士

일제시대 조선총독에게 폭탄을 투척, 대한 독립의 당위성을 세계에 알린 강우규(姜宇奎·1859~1920)의사가 11월의 독립운동가로 선정됐다고 국가보훈처는 31일 밝혔다.

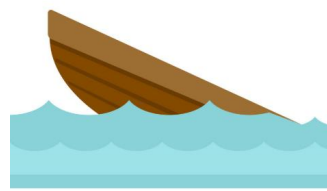
평남 덕천군에서 태어난 강 의사는 1883년 함남 흥원에서 사립학교와 교회를 세우 신학문 전파와 민족의식 고취에 앞장서다가 1911년 북간도로 망명했다. 1919년 3·1독립만세 운동이 일어나자 김립성 요하현에 자신이 설립한 광동학교에서 학생들과 동포들을 규합, 만세운동을 벌였다.

의사는 이어 조선총독이 사이토로 교체된다는 소식을 듣고 그를 저단기로 결심한 뒤 1919년 6월 국내로 잠입, 9월2일 환영식이 열리던 남대문역에서 사이토에게 폭탄을 던졌으나 폭탄에 실패했다.

의사는 재거사를 준비하다가 9월17일 체포, 이듬해 사형을 언도받고 서대문형무소에 수감됐다 교수형으로 순국했다. 정부는 1962년 의사의 공훈을 기려 건국훈장 대한민국장을 추서했다.

/廣龍源기자 kysu@chosun.com

Word Sense Disambiguation



조선일보

건강

항생제 개발로 '죽음의 감염병' 물리쳤다

머크 매뉴얼 선정 '20세기약학 10대업적'



다.그러나 이제 **의사**들은 병든 심장이나 간을 바꿔치기 할 수 있을 정도가 됐다.18
판이 발행된 '머크 매뉴얼(Merck Manual)'은 전세계 의사들이 가장 많이 참고하는

2판 제24515호 25

① 감염성 질환 극복

▲1940년대 이후 이른바 '3대 항생제'(살피계열 항생제, 페니실린, 스트렙토마이신)가 환자의 치료에 사용되면서 각종 감염질환의 사망률이 크게 떨어졌다. 한편의대 출신의 미생물학자 알렉산더 플레밍은 1928년, 우연히 보드 상구균의 발병을 억제하는 진균(페니실

리움)을 발견했으며, 1940년엔 영국에서 이를 안정된 분말로 정제하는데 성공했다. 이 약은 1944년 2차대전 말엔 군인을 치료하는데 최초로 사용되면서 항



소아마비-천연두 등 백신 개발로 자취 감춰
흑진단-치료 기술 발달... 초기 발견하면 원치도

병제 시대의 막을 열었다. 스트렙토마이신은 20세기 초 가장 흔한 사망원인이었던 결핵을 퇴치시키는 데 결정적인 공헌

을 했다. 진행된 이 역학조사할 통해 출연이 나 고함압과 같은 심장병 위험인자가 최초로 밝혀졌다.

AIDS 박제

11월 독립운동가 강우규 義士

일어나자 김립성 요하현에 자신이 설립한 광동학교에서 학생들과 동료들을 모아 마네편등을 꾸미다

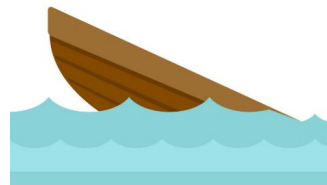
일제시대 조선총독에게 폭탄을 투척, 대한 독립의 당위성을 세계에 알린 강우규(姜宇奎·1859~1920) **의사**가 11월의 독립운동가로 선정됐다고 국가보훈처는 31일 밝혔다.

국민기립의 하기 위해 영화했다" 흥다터(고 3200여 의 시특시 스

1883년 함남 흥원에서 사립학교와 교회를 세워 신학문 전파와 민족의식 고취에 앞장서다가 1911년 북간도로 망명했다. 1919년 3·1 독립만세 운동이

순국했다. 정부는 1962년 의사의 공훈을 기려 건국훈장 대한민국장을 추서했다. /廣龍源기자 kysu@chosun.com

Word Sense Disambiguation





醫師

VS

義士

의사 'Eui- Sa'


의사 'Eui- Sa'

의사¹² 醫師 [의사]  ★★★ 

1. 명사 일정한 자격을 가지고 병을 고치는 것을 직업으로 하는 사람.
2. 명사 법률 서양 의술과 양약으로 병을 고치는 것을 직업으로 하는 사람.

유의어 의원³

표준국어대사전

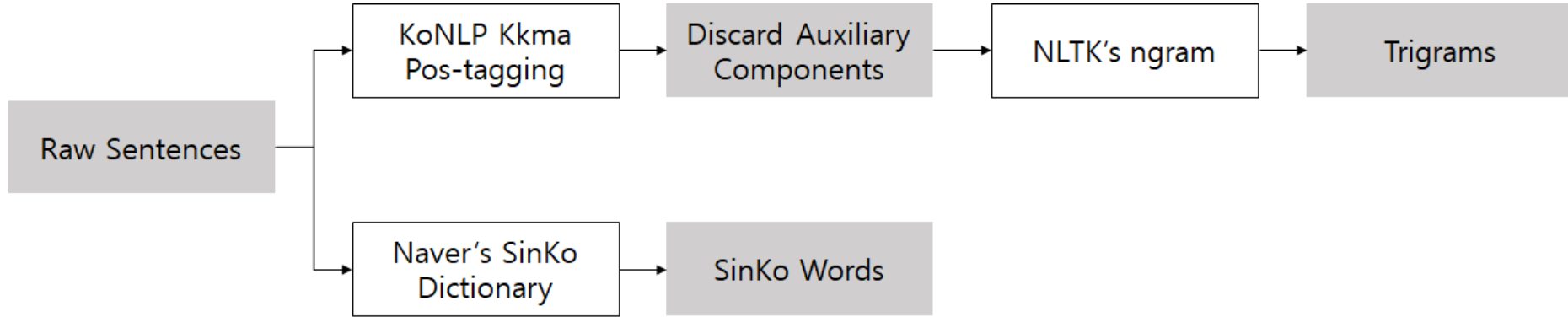
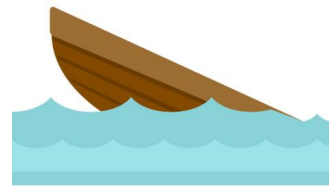
의사³ 義士 [의:사] 

명사 의로운 지사(志士).

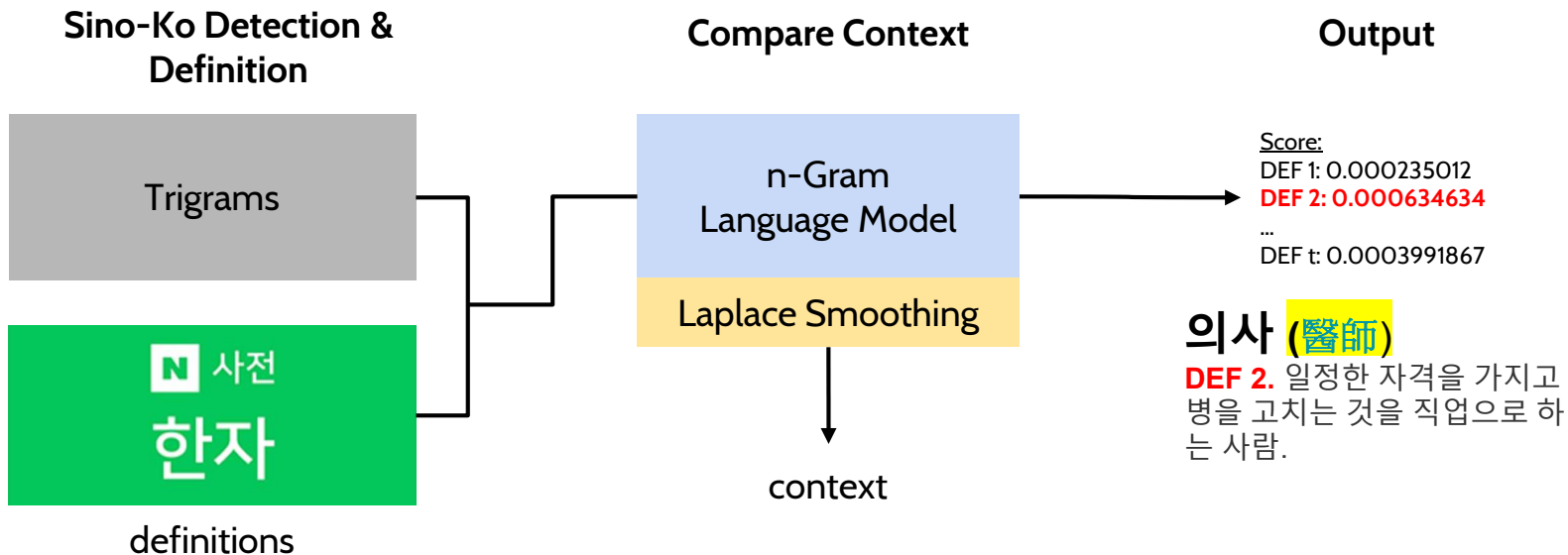
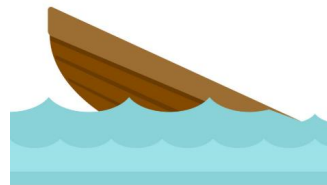
유의어 지사⁵ 열사¹

표준국어대사전



Word Sense Disambiguation

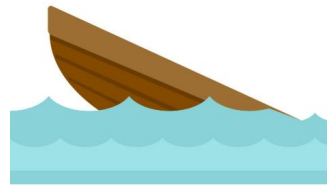


Word Sense Disambiguation



Demonstration/Results

-  Sino-Korean words
-  Difficult Sino-Korean words



오전 10시7분 증인선서 낭독으로 시작된 전 전대통령의 증언은 5공비리, 광주문제에 관한 답변 속에 저녁까지 계속됐다. 그러나 증언도중 증언내용이 부실하고 위증이 있다는 야당의원들의 고함과 의사진행 발언요구 등속에 소란이 끊임없이 벌어져 모두 7여 차례의 정회가 거듭되다. , 결국 밤 12시를 넘김으로써 전씨가 의사당을 떠나, 청문회는 전씨의 구두 답변을 다 듣지 못한 채 끝나고 말았다.

10.07am, the former president testimony, which began with reading the witness's oath, continued through the evening in response to the irregularities in the 5TH republic and Gwangju issues. However, during the testimony, there was a constant turmoil among the leading opposition party who said that the content of the testimony was poor and there was perjury. This resulted in 7 times of prorogation. Eventually, past 12:00, Mr. Jeon left the capitol, and the hearing ended without hearing all of his verbal responses.

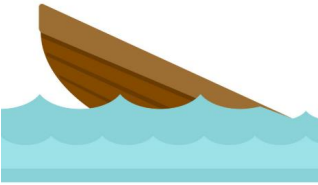
Demonstration/Results



Sino-Korean words

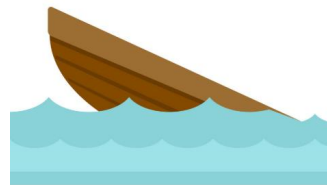



Difficult Sino-Korean words



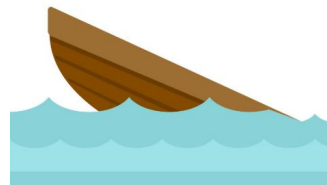
The screenshot shows a web browser window with a single tab titled "CS372 Team9". The address bar displays "localhost:3000". The browser's toolbar includes navigation arrows, a refresh button, a star icon, and several notification icons. The main content area has a dark blue background. Centered on this background is a white rectangular box containing the application's interface. At the top of this box, the text "SinK DaT" is displayed in a bold font, with "(Sino-Korean Detector and Translator)" underneath it. Below the title is a large, empty text input field with a light gray border and a small cursor icon at the bottom right. Underneath the input field is a blue button with the text "See results" in white. At the bottom of the white box, the text "Created by CS372 Team 9, 2020 Spring" is displayed. A mouse cursor is visible at the bottom right corner of the white box.

Evaluation



- In progress
- Three Core Modules: **Detect** → **Difficulty Measure** → **Translate** 
- **Detect / Translate:** Compare the system performance with human annotated result
- **Difficulty Measure:** Online survey with Korean Middle-School Students

Conclusion



- We have designed a **SinK-DaT model** that is able to detect and translate difficult Sino-Ko words in the newspaper corpus **with appropriate context**.
- We aim to help younger generation who are “weak” in Sino-Ko words to comprehend Korean literature!
- Efficient NLP tools such as POS-tagging and N-gram modeling were utilized.

Reducing the Generation Gap:

SinK DaT (Sino-Korean Detector and Translator)



Thank you 😊