

Mewnblaniadau Geiriau ar gyfer y Gymraeg

Geraint Palmer¹, Pdraig Corcoran², Laura Arman³, Dawn Knight³, ac Irena Spasic²

¹Yr Ysgol Mathemateg, Prifysgol Caerdydd

²Yr Ysgol Cyfrifiadureg a Gwybodeg, Prifysgol Caerdydd

³Yr Ysgol Saesneg, Cyfathrebu ac Athroniaeth, Prifysgol Caerdydd

Ariennir gan Lywodraeth Cynulliad Cymru

04/11/2020

Cynnwys

1. Beth yw mewnbaniadau geiriau?
2. Defnydd mewnbaniadau geiriau
3. Hyfforddi mewnbaniadau geiriau Cymraeg
4. Arddangos mewnbaniadau geiriau Cymraeg

Beth yw mewnblaniadau geiriau?

Diffiniad

Mapiad o'r gofod lecsio-semantig i ofod fector real.

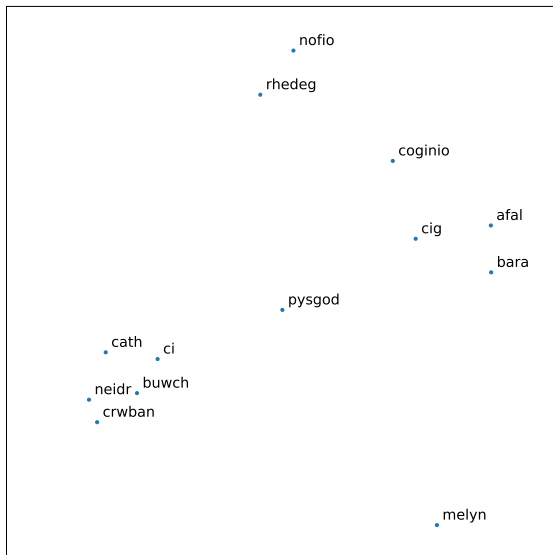
$$f(\text{cath}) = (24.5, 0.4, -8.4, \dots, -9.9)$$

$$f(\text{rhedeg}) = (-5.77, 12.0, 1.38, \dots, 32.6)$$

$$f(\text{melyn}) = (0.1, -0.8, -28.5, \dots, -2.7)$$

⋮

Priodweddau



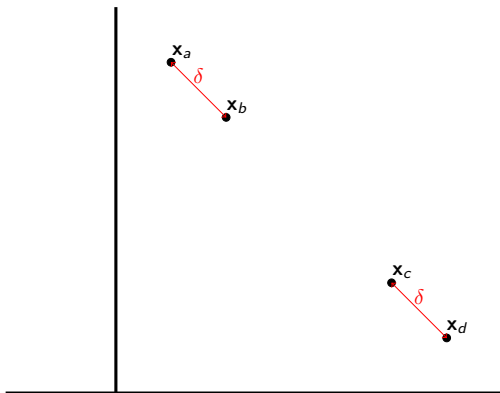
Priodweddau

$$f(\text{brenin}) = \mathbf{x}_a$$

$$f(\text{brenhines}) = \mathbf{x}_b$$

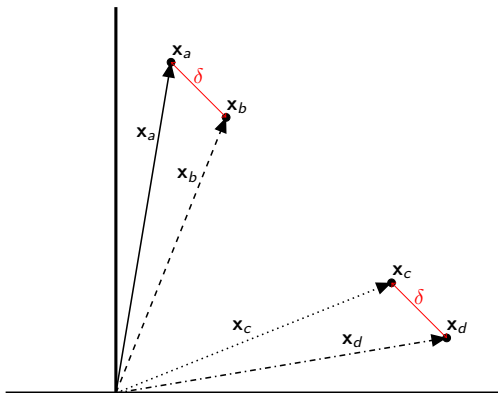
$$f(\text{myfyriwr}) = \mathbf{x}_c$$

$$f(\text{myfyrwraig}) = \mathbf{x}_d$$



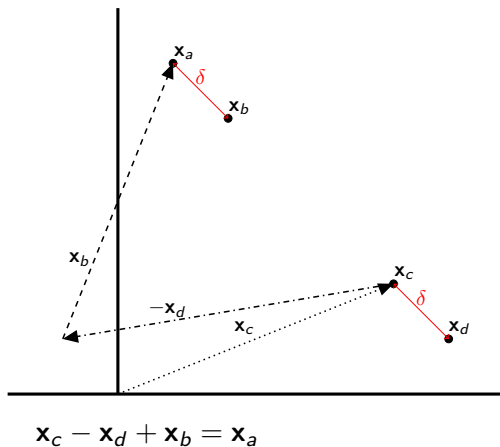
Priodweddau

$$\begin{aligned}f(\text{brenin}) &= \mathbf{x}_a \\f(\text{brenhines}) &= \mathbf{x}_b \\f(\text{myfyriwr}) &= \mathbf{x}_c \\f(\text{myfyrwraig}) &= \mathbf{x}_d\end{aligned}$$



Priodweddau

$f(\text{brenin}) = \mathbf{x}_a$
 $f(\text{brenhines}) = \mathbf{x}_b$
 $f(\text{myfyriwr}) = \mathbf{x}_c$
 $f(\text{myfyrwraig}) = \mathbf{x}_d$





Defnydd mewnbaniadau geiriau

Nifer mawr o modelau dysgu peirianyddol a phrosesu iaith naturiol, gan gynnwys:

- Cyfieithu peirianyddol
- Dadansoddi sentiment
- Adnabod endidau
- Parsio dibyniaethau


Defnydd - Cyfieithu peirianyddol


 “the distribution function is continuous due to this identity”


 “mae swyddogaeth y dosbarthiad yn barhaus oherwydd y hunaniaeth hon”

2018 Qi, Ye, *et al.* **When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?**
Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).

Defnydd - Cyfieithu peirianyddol





 “the distribution function is continuous due to this identity”

 “mae swyddogaeth y dosbarthiad yn barhaus oherwydd y hunaniaeth hon”

 “mae ffwythiant y dosraniad yn ddi-dor oherwydd yr unfathiant hwn”

2018 Qi, Ye, *et al.* **When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?**
Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).

Defnydd - Dadansoddi sentiment

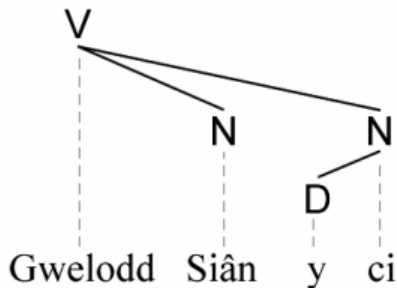
Sylw	Sentiment
“oni'n rili lico'r ffilm”	
“roedd stori'r ffilm yn sôn am long ofod”	
“y ffilm gwaethaf dwi 'di gweld erioed”	
“gor-actio bendigedig can Cameron Diaz”	

“darganfuwyd y prif weinidog bod
ganddi gancr y fron yr haf
yr oedd hi'n gweithio ar y papur
gwyn dadleuol hwnnw.”

Defnydd - Adnabod endidau

“darganfuwyd y prif weinidog ^{Person} bod
ganddi gancr y fron ^{Peth} yr haf ^{Amser}
yr oedd hi'n gweithio ar y papur
gwyn ^{Peth} dadleuol hwnnw.”

Defnydd - Parsio dibyniaethau



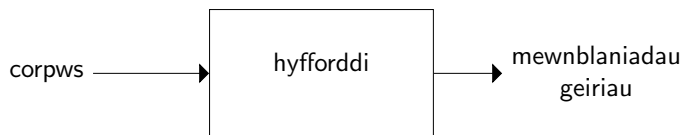
Darlun: 2019 Jones DB., Prys, D., Prys M. a Prys G., **Llwllyfr Technolegau Iaith**

2018 Dozat T. a Manning CD. **Simpler but More Accurate Semantic Dependency Parsing**. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Hyfforddi Mewnblaniadau Geiriau



Hyfforddi Mewnblaniadau Geiriau



← cyd-destun →

... aeth y ci i **mewn** i'r **tŷ** er mwyn **bwyta**'r bwyd i gyd ...

↑
gair

CBOW: mwyafsymio cyfanswm $\mathbb{P}(\text{cyd-destun} \mid \text{gair})$

Skip-gram: mwyafsymio cyfanswm $\mathbb{P}(\text{gair} \mid \text{cyd-destun})$

Casglu Corpws

<i>Ffynhonnell</i>	<i>Nifer o eiriau</i>
Beibl	749,573
Corwps An Crúbadán Corwps a chasglwyd yn UDA o blogiau, trydar a wefannau Cymraeg	22,572,066
CorCenCC Corwps rhagbaratoawl iaith electroneg CorCenCC	1,875,540
Cronfa Electroneg o Gymraeg Casgliad amrywiol o rhyddiaeth, nofelau, a dogfennau gweinyddol	1,046,800
Google Corpuscrawler Tecnolgydd Google i cael caffael ar corpera nifer o ieithoedd, yn bennaf BBC Cymru Fyw	14,791,835
Gwerddon	767,677
Project DECHE Prosiect Digidol, E-gyhoeddi a Chorpws Electronig, gwerslyfrau wedi'u digido	2,126,153
Trafodion Cynulliad Cenedlaethol 1999-2006	11,527,963
Trafodion Cynulliad Cenedlaethol 2007-2011	8,883,970
Wikipedia Cymraeg	21,233,177
Wefannau amrywiol Golwg360, O'r Peward Gwynt, Barn, a PoblCaerdydd	7,388,917
<i>Cyfanswm</i>	92,963,671

Arddangos Mewnblaniadau Geiriau Cymraeg

<https://datainnovation.cardiff.ac.uk/is/wecy/index.html>

Arddangos Mewnblaniadau Geiriau Cymraeg

ARIAN

- harian
- arian'
- ariannu
- cyllid
- bunnau
- gyllid
- bres
- arianu
- goffrau
- wario

Arddangos Mewnblaniadau Geiriau Cymraeg

CYSGU

- gysgu
- deffro
- llewygu
- cysgai
- chwtsho
- chysgu
- dihuno
- cerdded
- ddihuno
- gysgai

Arddangos Mewnblaniadau Geiriau Cymraeg

BLAWD

- blawdog
- flawd
- blawr
- blawdy
- siaradblawd
- menyn
- fenyn
- cyflasynnau
- blaw
- chwstard

Arddangos Mewnblaniadau Geiriau Cymraeg

TONYREFAIL

- donyrefail
- nhonyrefail
- ffosyrefail
- maesyrefail
- trebanog
- tonysguboriau
- tonteg
- tonypandy
- trecynon
- trealaw

Arddangos Mewnblaniadau Geiriau Cymraeg

ACTORES

- actor
- actoresau
- sgriptwraig
- digrifwraig
- berfformwraig
- comediwraig
- chantores
- gantores
- ddigrifwraig
- perfformwraig

- <https://datainnovation.cardiff.ac.uk/is/wecy/index.html>
- <http://www.geraintianpalmer.org.uk/>
- palmergil@cardiff.ac.uk
- @GeraintPalmer